Hybridisation in bluebells (*Hyacinthoides* spec.)

Using next-generation sequencing to reconstruct a natural hybrid zone in Spain

Jeannine Marquardt, MSc.[*]

December 2nd, 2016

Submitted in partial fulfilment of the requirements of the Degree of Doctor of Philosophy

Supervisors:
Prof Andrew R Leitch
Prof Richard A Nichols
Prof Harald Schneider (external)

Queen Mary University of London

[*] jeannine.marquardt@qmul.ac.uk

Statement of Originality

I, *Jeannine Marquardt*, confirm that the research included within this thesis is my own work or that where it has been carried out in collaboration with, or supported by others, that this is duly acknowledged below and my contribution indicated. Previously published material is also acknowledged below.

I attest that I have exercised reasonable care to ensure that the work is original, and does not to the best of my knowledge break any UK law, infringe any third party's copyright or other Intellectual Property Right, or contain any confidential material.

I accept that the College has the right to use plagiarism detection software to check the electronic version of the thesis.

I confirm that this thesis has not been previously submitted for the award of a degree by this or any other university.

Signature:

Date: December 2nd, 2016

Abstract

Hybridisation is a common evolutionary process that can arise in primary or secondary contact. Gene flow and/or reproductive isolation between hybridising taxa can be explored in hybrid zones. Therefore, a (homoploid) hybrid zone in north-west Spain between *Hyacinthoides non-scripta* and *H. hispanica* was studied. The centre occurs west to east across the Galicio-Duero Mountains with *H. non-scripta* distributed north, and *H. hispanica* south of the centre. The hybrids' genome sizes and phenotypes represented a range of intermediate states between their parents. Crossing and seed germination experiments revealed a low inter-specific barrier, and the hybrids showed similarly good fitness. Genome wide markers for large genome species were designed from transcriptomes. Diagnostic SNPs between *H. non-scripta* and *H. hispanica* were targeted and re-sequenced with multiplexing PCR. Coalescence analyses suggested a Pleistocene origin of parapatric speciation between *H. non-scripta* and *H. hispanica*. These results are supported by shared inter-specific polymorphisms, the lack of recent hybrid generations and of parental individuals in sympatry. Differential introgression patterns between the organellar and nuclear genomes revealed that formerly *H. hispanica* ranged further north but was swamped by *H. non-scripta* alleles. Asymmetric hybridisation was reasoned by absence of backcrosses between northern hybrids to *H. non-scripta*, but presence between southern hybrids and *H. hispanica*. Combining these results, a southwards movement of the hybrid zone centre caused by climate change (and adaptive introgression), or inter-specific differences in flowering time was suggested. Cline patterns revealed cyto-nuclear incompatibilities, which could evolve through divergent adaptation of the organelle to climate and a delayed selection on nuclear inter-acting loci. Both species are in secondary contact in the UK due to recent introduction(s) of *H. hispanica* and garden variants, which is considered to cause genetic pollution of native *H. non-scripta*. Therefore, a conservation study is in progress, in which this diagnostic marker system for bluebells is applied.

Contents

List of Figures

List of Tables

Acknowledgement

This work was supported by a travel grant from the QMUL Doctoral College and the European Marie Curie Initial Training Network 'INTERCROSSING' [PITN-GA-2011-289974].

I would like to express my thanks to Steve W Ansell, who provided key data to this project allowing the investigation of the Spanish part of bluebell hybridisation research. Further, I appreciated the support and opportunity to use the excellent facilities at the Natural History Museum and the Royal Botanic Gardens, Kew in London. Thanks to Professor Harald Schneider, and Professor Dirk Metzler for your long-term professional friendship. Special thanks reach out to Professor Andrew R Leitch and his great team of students in 2013-2015. I valued the scientific discussions and advice from the whole of the SBCS department throughout my graduate time at Queen Mary University of London. Thanks to all early career smart-ass students and their PIs of the INTERCROSSING network; and of course Professor Richard A Nichols for bringing us all together. Most of all, Alexandre, thank you for our enjoyable collaboration and your great help and contributions. I appreciate the collegial correspondence and support from Deborah Cohn, Markus Ruhsam, and Pete Hollingsworth from the Royal Botanic Garden Edinburgh, and Professor Joachim Hermisson and his group of 2015/2016, especially during my time in Vienna.

I am grateful for the opportunity of this endeavour and my special thanks go out to those who supported me during this journey: by listening, by ranting on, and by keeping my spirits up. The list of friends, people and acquaintances is long – you know who you are! Thank you ever so much. My gratitude is further devoted to my family for guidance and support in the moments of making decisions and defeating doubts.

The Blue Bell

Emily Brontë (1818-1848)

The blue bell is the sweetest flower
That waves in summer air;
Its blossoms have the mightiest power
To soothe my spirit's care.

There is a spell in purple heath
Too wildly, sadly dear;
The violet has a fragrant breath
But fragrance will not cheer.

The trees are bare, the sun is cold;
And seldom, seldom seen;
The heavens have lost their zone of gold
The earth its robe of green;

And ice upon the glancing stream
Has cast its sombre shade
And distant hills and valleys seem
In frozen mist arrayed -

The blue bell cannot charm me now
The heath has lost its bloom,
The violets in the glen below
They yield no sweet perfume.

But though I mourn the heather-bell
'Tis better far, away;
I know how fast my tears would swell
To see it smile today;

And that wood flower that hides so shy
Beneath the mossy stone
Its balmy scent and dewy eye:
'Tis not for them I moan.

It is the slight and stately stem,
The blossom's silvery blue,
The buds hid like a sapphire gem
In sheaths of emerald hue.

'Tis these that breathe upon my heart
A calm and softening spell
That if it makes the tear-drop start
Has power to soothe as well.

For these I weep, so long divided
Through winter's dreary day,
In longing weep--but most when guided
On withered banks to stray.

If chilly then the light should fall
Adown the dreary sky
And gild the dank and darkened wall
With transient brilliancy,

How do I yearn, how do I pine
For the time of flowers to come,
And turn me from that fading shine
To mourn the fields of home –

List of Abbreviations

ABC – approximate Bayesian computation

AF – allele frequency

APG – Angiosperm Phylogeny Group

biSNP – bi-allelic single nucleotide polymorphism

BM – Plant collection catalogue of the Natural History Museum

cal B.P. – radiocarbon calibrated time point before 1950

cds – coding sequence

CI – confidence interval

cor – correlation

cov – coverage

CPG – Chelsea Physics Garden London

CV – coefficient of variation

df – degree of freedom

FCM – flow cytometry

Gb – giga bases (the genome size measure)

GLMM – generalised linear mixed-effect model

GT – genotype

HWE – Hardy-Weinberg equlibrium

jsfs – (folded) joint site frequency spectrum

ka – thousand years ago

mya – million years ago

ORF - open reading frame

PC – principal component

PE – paired end

pg – pico gram (the genome size measure)

RBGE – Royal Botanic Garden Edinburgh

s.l. – *sensu lato*

SD – standard deviation

SNP – single nucleotide polymorphism

Chapter 1

Introduction

1.1 Evolutionary history of bluebells

The genus *Hyacinthoides* Heist. ex Fabr. includes about eleven species, which are bulbous perennials. Their taxonomy and relationships were most recently phylogenetically revised by Grundmann et al. (2010). A clade containing these species is placed into subfamily Hyacinthoideae of Hyacinthaceae (Pfosser and Speta, 1999); although the family has been included, more recently, in an expanded Asparagaceae *s.l.* (APG III, 2009) as subfamily Scilloideae (Chase et al., 2009). Within this thesis, the taxonomic ranks are based on Wetschnig and Pfosser (2003) and Pfosser et al. (2003). The APGII (APG II, 2003) or APGIII (APG III, 2009) classification are highlighted when they have been used as the classification system in a cited reference.

Asparagales includes about 50 % of the monocot species diversity, but relationships between the major lineages are still only partly resolved (Chen et al., 2013). The family Hyacinthaceae *sensu* Pfosser and Speta (1999) comprises approximately 1,000 species grouped in 35 genera (Buerki et al., 2012). Molecular studies identified four monophyletic groups within Hyacinthaceae (Manning et al., 2003; Pfosser and Speta, 1999): Hyacinthoideae, Ornithogaloideae, Urgineoideae and Oziroeoideae. These subfamilial ranks were decreased to tribes in Chase et al. (2009) and Haston et al. (2009). Their biogeography has been tackled using approaches such as dated phylogenies of the family Hyacinthaceae (Buerki et al., 2012) or of subfamily Ornithogaloideae (Martinez-Azorin et al., 2011); inferences of ancestral biogeography were proposed for the subfamilies Hyacinthoideae and Urgineoideae (Ali et al., 2012, 2013).

The subfamily Hyacinthoideae (or tribe Hyacintheae *sensu* Chase et al. (2009)) has been traced back to a sub-Saharan-African origin during the Early Miocene (Ali et al., 2012; Buerki et al., 2012). From there its members either dispersed frequently before 20 million years ago (mya) (Buerki et al., 2012), or from a single event between 19 and 18 mya (Ali et al., 2012) to the Mediterranean and eastern Asian region. The monophyletic tribe Hyacintheae (*sensu* Pfosser et al., 2003; Wetschnig and Pfosser, 2003) is restricted to Eurasia; however, the dispersal route to Asia may have been possible through Europe or from the Mediterranean region (Ali et al., 2012).

The bluebell genus *Hyacinthoides* has a strictly Mediterranean origin with radiations to northern Europe (node 63 of Fig. 1 in Ali et al., 2012). It is divided phylogenetically

into three monophyletic groups, with a deep split between a western and eastern clade. This split is supported by differentiated flowering times, with north African taxa flowering in winter-autumn and Iberian taxa flowering in spring (Fig. 1 in Grundmann et al., 2010). Within the eastern clade, *H. italica* shows a distinct range that is restricted to the Maritime Alps of France and Italy whilst the other three species (*H. aristidis*, *H. ciliolata*, and *H. lingulata*) occur in Northern Africa (Grundmann et al., 2010). The western clade contains two monophyletic groups and the species distributions overlap in the southern Iberian Peninsula and north-west Morocco. The divergence within the clade including *H. mauritanica*, *H. flahaultiana*, and *H. reverchonii* was suggested to be a consequence of speciation by geographic isolation, since there is no range overlap in current species distributions (W1 in Fig. 1 in Grundmann et al., 2010). The second western clade comprises the widespread and most northerly bluebell species, *H. non-scripta*, *H. hispanica*, which is restricted to the Iberian Peninsula, *H. paivae*, an endemic species of north-western Spain, and the tetraploid *H. cedretorum*, occurring in wider ranges of northern Africa (W2 in Fig. 1 in Grundmann et al., 2010). The *trnL-F* plastid markers used for phylogeographic studies, however, resulted in only two haplotypes, one of *H. non-scripta* and one of *H. hispanica*. The latter is also found in *H. paivae* and *H. cedretorum*, although they occur outside of *H. hispanica*'s range. Moreover, *H. cedretorum* is either of autopolyploid origin from *H. hispanica*, or an allopolyploid with an unknown parent. The phylogenetic position of *H. paivae* remains unresolved and its distinct flower shape continues to puzzle researchers (Ortiz and Rodríguez-Oubiña, 1996; Ortiz et al., 1999).

Molecular clock dating using plastid sequences suggested an origin of the genus about 5.81 (mya) (more specifically only *H. non-scripta* and *H. italica* were included, Ali et al., 2012), while older estimates range from 30 - 15 mya (Buerki et al., 2012) with an estimate for the *H. non-scripta*-*H. hispanica* clade from 15 - 5 mya. The deeper estimate for the diversification of the genus *Hyacinthoides* was postulated to have been influenced by the opening of the Strait of Gibraltar in the Early Pliocene, causing vicariance. Alternatively, the clade may have diversified as a consequence of large-scale desiccation of the Mediterranean during the Messinian salinity crisis (around 5 mya, Grundmann et al., 2010). More recent diversification events, for example the split between *H. non-scripta* and the clade including *H. hispanica*, are thought to have been caused by Pleistocene glaciation cycles (since 2.58 mya; Grundmann et al., 2010).

The current species range of the British bluebell, *H. non-scripta* (L.) Chouard ex Rothm. expands from northern Spain along the Atlantic coast to the British Isles with about 25 - 50 % of the British bluebell's population present in the British Isles (Ingrouille, 1995). In contrast, *H. hispanica* (Mill.) Rothm., naturally occurs in the west to central regions of the Iberian Peninsula (Grundmann et al., 2010). Seed germination in bluebells (*H. non-scripta*, and likely also *H. hispanica*) is restricted to a two-phase temperature treatment, which is adapted to their distribution range (Blackman and Rutter, 1954; Thompson and Cox, 1978; Vandelook and van Assche, 2008). Their distribution is likely confined by the highest temperature in summer (and consequently by drought) and the lowest temperature in winter (Blackman and Rutter, 1954; Walter and Hengeveld, 2000). Grundmann et al. (2010) postulated that *H. non-scripta* expanded post-glacially from northern Iberia to its present species range including Scotland, UK. After the last glacial

maximum the re-colonisation of the British Isles from Europe was only possible through land bridges within a relatively short period before the British Isles were separated by the North Sea and the English Channel (at about 8,000 ya, Hewitt, 1999). Temperate tree species from their northern-most refugia along the 46°N would have required a migration rate of at least 50 m/year (Svenning and Skov, 2007a). However, using a transplant experiment and observations after 45 years, a dispersal distance for *H. non-scripta* of 0.6 – 6 cm/year was estimated (van der Veken et al., 2007). This means that within c. 12,000 years *H. non-scripta* would have naturally expanded its range by only 720 m and would have been unable to reach the British Isles from refugia in northern parts of the Iberian Peninsula. One could speculate that the dispersal rate of bluebells by seeds has been underestimated by neglecting rare long range events, that there could have been post-glacial expansion from northern refugia, in for example Southern France (Svenning and Skov, 2007b), or that there was plant dispersal by humans in older and more recent times (Hodkinson and Thompson, 1997). For example, bluebells are very rich in secondary compounds that could find ethnomedicinal applications. Fructan is stored in the bulb as the principal reserve carbohydrate and sucrose as the second reserve carbohydrate in the shoot (Brocklebank and Hendry, 1989). The sap of bluebells is gooey, and anecdotal evidence suggests that the sap has been used to stick feathers to arrow-heads in the Bronze age, to stick papers by bookbinders, and that the Elizabethans used the starch of bluebell juice to stiffen sleeves and collars (Simmonds, 2004, and references therein). Further, bluebells contain glycosidase-inhibiting alkaloids (i.e. nitrogen analogues of mono- and disaccharides; Watson et al. (1997)) and other alkaloids (Kato et al., 1999) that cause abdominal pain, dysentery, lethargy and dullness in mammals that have eaten bluebells (Simmonds, 2004; Watson et al., 1997). Supposedly, bluebells have been used to treat leprosy, snake bites (Simmonds, 2004), and leucorrhoea (discharge of mucus from the vagina)[1]. Such compounds are not just restricted to *Hyacinthoides*, and molecule presences/absences and molecule structures provide taxonomically relevant information for the Hyacinthaceae *sensu* APG II (2003) (Mulholland et al., 2013). Other members of the Hyacinthaceae have also been long used in traditional medicine (Mulholland et al., 2013). However, apart from these reports not much is known about the medicinal uses of bluebells in British folklore, nor are there products such as cosmetics containing an extract of bluebells (Thoss et al., 2012). The most recent investigations have explored the possibility of commercial use of bluebell's seed oil, which is 'sufficiently unusual to generate interest in the chemical exploitation' (Thoss et al., 2012).

1.2 Hybridisation and conservation interest of the British bluebell

Hyacinthoides non-scripta has iconic status in the British Isles (Kohn et al., 2009; Thoss et al., 2012). It is well known from ancient woods by its remarkably dense occurrences of blue-violet flowers and a heavy scent dominating the field-layer in early spring (Pigott, 1984). Indeed, *Hyacinthoides non-scripta* was suggested as an indicator species of ancient (British) woodlands (Rose, 1999), and are a focus of nature conservation due

[1] http://www.kew.org/science-conservation/plants-fungi/hyacinthoides-non-scripta-bluebell

to their occurrence in supposedly primary forests (Goldberg et al., 2007). In France and southern Belgium too, the conservation priority of woodlands with dense bluebell cover has been highlighted (Pigott, 1984). Bluebells (likely applicable to all congenerics) are vulnerable to grazing, cutting, or trampling because the number of leaves is pre-defined in the previous year and cannot be renewed in the current season when destroyed, leading to a resource deficit in the present and subsequent year (Cooke, 1997; Grime et al., 1988; Sims et al., 2014). In 1998, *H. non-scripta* was included as specially protected under the UK Wildlife and Countryside Act from 1981[2] due to commercial over-exploitation (Kohn et al., 2009). Additionally, for more than a decade, hybridisation has been observed between the British bluebell and presumably *H. hispanica* in the British Isles. The invasion of human-introduced bluebell garden varieties and *H. hispanica* raised concerns, as introgressive hybridisation could pose a threat of 'genetic pollution' to the native *H. non-scripta* (Kohn et al., 2009). Current research is under way to explore the invasiveness of such 'alien' taxa in the UK, including censuses by societies (Plantlife International: Dines, 2005; Pilgrim and Hutchinson, 2004) and, more recently, citizen science projects by various research institutes and societies, for instance the Natural History Museum London in collaboration with the Royal Botanic Garden Edinburgh[3], and the Wellcome Genome Campus Public Engagement team joint forces with the Eden Project and The Wildlife Trust[4]. The latter project applies genetic DNA barcoding (of plastid markers) in addition to locality information using a mobile application[5].

These approaches have their scientific merits and value for conservation management (e.g. Ballard et al., 2016), and most importantly they raise public awareness of alien bluebells, which are escaping urban areas. However, morphological census studies might underestimate the 'genetic pollution' by introgressive hybridisation and breakdown of species identity. For instance, in the 'NHM London Bluebell Survey' the species identification relies on field pictures and categorical morphological descriptions of three taxa: parent one, parent two, and hybrids. This assessment is in conflict with the known presence of additional garden cultivars with different phenotypes (Grundmann et al., 2010; Kohn et al., 2009; Page, 1987). The segregation of parental phenotypes within bluebell hybrids has not been studied. To exemplify the problems, triploid individuals were identified as pure samples of *H. non-scripta* and *H. hispanica* (supplement of Grundmann et al., 2010), although polyploid taxa are rare within *Hyacinthoides* with one tetraploid exception (*H. cedretorum*). Polyploidy due to hybridisation has been observed in a number of cases, although triploid hybrids mostly resulted from diploid and tetraploid crosses (Hanusova et al., 2014; Lowe and Abbott, 2015; Prancl et al., 2014). Additionally, the taxon *H. hispanica* from the Iberian Peninsula and the deemed invasive plants of the so-called '*Spanish bluebell*' in the UK could potentially be from markedly different genotypes, accessions, or cultivars (Grundmann et al., 2010; Page, 1987).

[2] www.jncc.gov.uk/page-3614; accessed Nov-25–2016

[3] www.nhm.ac.uk/take-part/citizen-science/bluebell-survey.html; accessed Nov-25–2016

[4] publicengagement.wellcomegenomecampus.org/page/bluebell-survey-2016; accessed Nov-25–2016

[5] plus.epicollect.net/blue bells/Bluebells; accessed Nov-25–2016

1.3 Aims and objectives of this thesis

What previous projects are missing is a genome-wide genetic marker set that can tell apart all alien taxa, hybrids, and the native British bluebell, which are closely related. Such data could also be applied for a population genetics study to determine the extent of introgressive hybridisation from urban areas into ancient bluebell woodlands. In addition, there is a lack of knowledge about the reproductive isolation between *H. non-scripta* and *H. hispanica*, the frequency of hybrid formation, the segregation of parental morphologies in hybrids, and the origin of triploid samples.

Hybridisation between *H. non-scripta* and *H. hispanica* has been discovered away from the British Isles, where both species' ranges naturally adjoin in northern Spain (Grundmann et al., 2010). Along the Cantabrian Mountains a few sites of mixed allozyme genotypes and plastid haplotypes of either parental species were discovered (Ansell et al., In prep.). Such a natural hybrid zone – called the bluebell hybrid zone has a seemingly latitudinal gradient from *H. non-scripta* (north) to *H. hispanica* (south). If this hybrid zone is not a mosaic hybrid zone, then it can be used to clearly define parental populations and explore the transition of characters, including morphology, genome size, and alleles.

For this thesis, a thorough sampling of individuals was conducted across the bluebell hybrid zone from the Galicio-Duero Mountains (where the hybrid zone has been reported; see chapter 2). Sufficient DNA material for a population genetics study across the hybrid zone and living plant material (bulbs) was collected. The bulbs were maintained over the course of the PhD in the UK and used for morphological characterisation of hybrids and genome size measurements. Lastly, crossing and germination experiments were used to explore the breeding system of the inter-acting species and their hybrids, in addition to the frequency of F_1 hybrid formation.

For species with large genomes (including bluebells, Bennett, 1972), limited genomic resources are available because of their inherent genomic complexity, and no species with a large genome is used as model organism in genetic studies (Feuillet et al., 2011; Michael and Jackson, 2013). For instance, no whole genome sequencing is available for any species that is closely related to the genus *Hyacinthoides* (Michael and Jackson, 2013). Nevertheless, species with large genomes can be more threatened by extinction (Vinogradov, 2003) and consequently there is demand for evolutionary or population genetics studies in such species, despite their inaccessible large genome. To analyse large genomes by DNA sequencing, reduction or enrichment are needed to reduce the complexity of the task (reviewed in Cronn et al., 2012). Using the transcribed portion of the genome (mRNAs) can be particularly suitable because there is often only little difference in the number of genes between species (Michael and Jackson, 2013). However, for a population genetic study that requires field sampling, RNA from tissues is not convenient as it is sensitive to quick RNA degradation. Consequently, here, transcriptome-based enrichment was applied to a few individuals that was then used to develop diagnostic markers for targeted amplicon re-sequencing from DNA across multiple individuals and populations (e.g. Guo et al., 2015; Salgado et al., 2014; Vatanparast et al., 2016). Diagnostic markers can be especially informative across a hybrid zone as they are fixed for different alleles in the hybridising parental populations, and then segregate in hybrids as heterozygous loci (Barton and Gale, 1993; Walsh et al., 2016). Therefore, in this thesis, a novel bioinformatics

approach was developed in chapter 3 to obtain hundreds of genomic markers, which are specifically designed to study allele frequencies across the bluebell hybrid zone.

The aim of chapter 4 is to understand the drivers of hybridisation between *H. hispanica* and *H. non-scripta* under natural conditions without the influence of humans on their dispersal capabilities. The developed marker set was applied to hundreds of samples, which had been collected from the hybrid zone area. The number of markers with diagnostic allele frequencies was assessed, along with their utility for detecting signals of population structure and genetic differentiation. Cline analysis was used to determine the cline centre positions. The distribution of single nucleotide polymorphisms (SNP) can be indicative of reproductive isolation. Varying patterns and shapes of SNP clines could provide information about the extent and direction of introgression (Barton and Gale, 1993). Finally, the same data set was used to perform coalescence simulations of different evolutionary models. The initial hypothesis is that both species met as secondary contact following the last glacial maximum. Different scenarios were used to model symmetric or asymmetric migration between the parental populations into the hybrid population. Alternative models, such as parapatric speciation between both species were also tested.

Chapter 2

Intermediate hybrids between *Hyacinthoides non-scripta* and *H. hispanica* – evidence from non-genetic data

2.1 Introduction

The British bluebell, *Hyacinthoides non-scripta* (L.) Chouard ex Rothm., has iconic status in the British Isles (Kohn et al., 2009; Thoss et al., 2012), and naturally occurs throughout the British Isles and along the Atlantic coast from the Netherlands to Northern Spain (Grundmann et al., 2010; Turrill, 1952). It is the only species of eleven congeneric species that has spread into northern Europe in post-glacial times (Grundmann et al., 2010). In the last decade, conservation concerns were raised because *H. non-scripta* was frequently found to hybridise with *H. hispanica* (Mill.) Rothm. in the British Isles (Dines, 2005; Pilgrim and Hutchinson, 2004). Although, the identity of the '*Spanish bluebell*' as treated in British floras (Stace, 1997) with *H. hispanica* is questionable based on morphological comparisons with individuals from its native range (Grundmann et al., 2010). Therefore, it is unclear if the '*Spanish bluebell*' that was first introduced as an ornamental plant in 1683 (Kohn et al., 2009; Pilgrim and Hutchinson, 2004) was a true representative of *H. hispanica* from the Iberian Peninsula, or was already a distinct variety or cultivar back then. The hybrids were first recognised in the wild in 1963 and have since increased in abundance along with the '*Spanish bluebell*' (Preston et al., 2002). In addition, triploid bluebells have been found in the UK (Grundmann et al., 2010). Therefore, this alien taxon can be considered as an invasive of ancient woodlands, which might be outcompeting the native British bluebell in its habitat (Kohn et al., 2009; Langdon, 2007; Pilgrim and Hutchinson, 2004). Other fears are that introgressive hybridisation can lead to a dilution of the British bluebell's gene pool with potential harmful effects – so-called genetic pollution (Stout, 2011; Todesco et al., 2016).

Hybridisation is a common evolutionary process in plants (Whitney et al., 2010) and animals (Schwenk et al., 2008). Increasingly, the consequences of climate change, human environmental impacts, and human introductions are causing the movement of species

into new habitats, where they genetically compete with the local diversity (Brennan et al., 2015; Buhk and Thielsch, 2015; Chown et al., 2015; Gómez and Lunt, 2006; Schierenbeck and Ellstrand, 2009; Todesco et al., 2016). Hybrid zone studies, which trace changes in traits over a small geographic scale (kilometres) with high resolution, present 'evolutionary laboratories' (Hewitt, 1988) that can be used 1) to study the strength of reproductive isolation between hybridising taxa (Charlesworth and Charlesworth, 2000; Lafon-Placette et al., 2016), 2) to identify genomic regions that are driven by adaptive introgression (Arnold and Martin, 2009; Martinsen et al., 2001), and 3) to determine environmental (biotic or abiotic) factors that promote or limit hybridisation (e.g. Hamilton et al., 2014). For bluebells in the UK, the probability of introgression between native and alien bluebells has been indicated by plant census recording of putative hybrids near forests that are the habitat of native bluebells (Kohn et al., 2009; Pilgrim and Hutchinson, 2004). These hybrids could have consequences for bluebell conservation and management, but it is unknown how strong are their capacities for introgression into native British bluebell populations.

In the central sierras of Spain, the species are parapatric (Grundmann et al., 2010) and the plants in this area represent a natural laboratory to study the dynamics of introgressive hybridisation. *Hyacinthoides hispanica* is bounded to the North at the southern end of the Cantabrian Mountain range (Grundmann et al., 2010), while *H. non-scripta*occurs in montane regions of the Cantabrian Mountains and partly at the northwestern Atlantic coast of Spain (Grundmann et al., 2010).

The genus *Hyacinthoides* includes species that are bulbous perennials, which can have one to several racemes per bulb with each raceme producing seven to 20 flowers (Corbet, 1998). Flower form can be used to distinguish *H. non-scripta* and *H. hispanica* morphologically. Flowers of both species are monoecious and naturally pollinated by insects (Blackman and Rutter, 1954; Kohn et al., 2009; Willmer, 2011). *Hyacinthoides non-scripta* has narrow tubular flowers with cream-coloured pollen. The flowers occur unilaterally on the raceme, which bends down slightly at the tip. The perianth segments are distinctively curled back at the end and slightly fused at the base (Deroin, 2014). In contrast, *H. hispanica* has bell-shaped, more open flowers with an intense blue anther colour. The raceme is erect and the flowers point to all sides. The perianth itself has a paler colour compared to the blue of *H. non-scripta*. These characters were identified from natural collections (Grundmann et al., 2010; Ortiz et al., 1999).

In this study, the natural hybrid zone between *H. non-scripta* and *H. hispanica* in Northern Spain and their hybrids were studied cytogenetically and morphologically. In addition, experimental crosses were conducted to determine the likelihood of hybrid formation and the fitness of hybrids.

2.2 Material and Methods

2.2.1 Study site and data collection

Fieldwork was conducted in the eastern part of the province Galicia and western part of the province Castile and León in Spain over two weeks in April to May 2013 (Figure 2.2). The aim was to visit equal numbers of collecting sites for each parental

species, *H. non-scripta* and *H. hispanica*, and their hybrids. The two species and their hybrids were identified using the identification criteria defined in recent taxonomic treatments (Grundmann et al., 2010), especially flower shape and pollen colour as described above. In the centre of the hybrid zone, the pairwise distance between collecting sites ranged from 0.6 – 63.5 km (mean 19.4 km). However, towards the edges of the parental ranges, fewer sites were found and therefore the distances between collecting sites were larger. Consequently, *H. non-scripta* was collected over a range of 2.7 – 153 km pairwise distances (mean 46.8 km) and for *H. hispanica* the distances ranged from 4.3 – 89.4 km (mean 38 km). At each collecting site, leaf tissue of at least ten different individuals was collected about ten meters apart (if possible), and at least three living bulbs were dug up for the cultivation experiments. In addition, the approximate population size, habitat, associated plant community, locality information, GPS coordinates and altitude in the field were recorded (Table A.2). The leaf material was dried and maintained in silica gel. Subsequently to fieldwork (and flow cytometry), the bulbs were potted individually in a shaded zone designated for plant cultivation at the roof of the Fogg building, Queen Mary University of London. The soil was mixed from peat-free multipurpose soil, vermiculate, perlite and coconut fibre (ratio 6:1:1:2 respectively). These maintained bulbs were used for hand-cross pollination experiments and morphological scoring in the subsequent years.

Genome size measurements were made from these plants collected in Spain, and in addition from fresh leaf material collected from eight bluebell accessions, obtained in May 2013 from Chelsea Physics Garden (CPG), London UK. Colleagues from the Natural History Museum London collected these samples from the Iberian Peninsula in 2008 and maintained them in the CPG (Table A.1, accessions). For more details on methodology of flow cytometry and chromosome squashes see section 2.2.4.

2.2.2 Morphological scoring

Observations during the fieldwork indicated a phenotypic gradient in which hybrids are more similar to their closest parental species than to the more spatially distant species. To quantify this hypothesis five morphological characteristics of bluebell plants were scored. The scores were obtained by combining observations made from the cultivated individuals during their blooming season from April to May in the years from 2014 to 2016 with data preserved in images taken during the fieldwork. The five characteristics were: 1) the habit of the plant; 2) the shape of the flower dominated by the perianth; 3) the degree of pigmentation in epidermal cells of the mature but still unopened anther, and 4) the colour when pollen was dehisced; and lastly 5) the pollen grain colour that probably also included some pollen kit. See Table 2.1 for more details. A principal components analysis was done in R version 3.3.1 (stats-package; R Core Team, 2016) using an Euclidean distance within each trait and the trait variables were scaled to have a unit variance. Samples with missing information for any of the five characteristics were removed from the analysis.

Table 2.1 – Overview of keys used in morphological scoring of plants and their values.

Trait	Key	Parameter range
Inflorescence	'habit'	1 (one-sided) – 3 (erect)
Shape of flower	'perianth_width'	1 (narrow-tubular) – 3 (wide open)
Anther epidermis colour before dehiscence	'anther_col'	1 (no pigmentation) – 5 (dark green)
Anther epidermis colour after pollen dehisced	'epi_pigm'	1 (no pigmentation) – 5 (dark green)
Pollen grain colour	'pollen_col'	1 (white) – 5 (very dark blue)

2.2.3 Experimental design of hand-cross pollinations

Three hand cross-pollination experiments were performed over two subsequent seasons (Figure 2.1). In 2014, one experiment was performed, in which individuals from all taxa were used to self-pollinate and to pollinate other plants within a taxonomic unit, i.e. *H. hispanica*, *H. non-scripta*, northern hybrids, and southern hybrids. In 2015, two experiments were performed simultaneously, one with plants of the parental species (i.e. *H. non-scripta* and *H. hispanica*), and the other with plants identified as hybrids in the field. Using crossing experiments, several questions regarding breeding system and likelihood of hybridisation between bluebells were addressed:

1. Are hybrids and *H. hispanica* self-incompatible, as reported for *H. non-scripta*?
2. Is the frequency of F_1 hybrids asymmetrical as a result of a preferred role of one parent as ovule donor or one parent as pollen donor?
3. How well can hybrids produce seeds, and does the geographic distance to pollen donor influence the seed set (i.e. in- or outbreeders)?

General scheme of the hand cross-pollinations. The plants were kept under laboratory (2014, climate control of 12/12h dark/light phases at room temperature around 21°C) or greenhouse conditions (2015, no direct sunlight and closed windows to avoid accidental pollinators). For each experiment, plants that produced flowers were selected and assigned to be either pollen donor or pollen receiver. During the flowering season each morning the anthers of pollen receivers were removed prior to flower opening and before matured pollen dehisced to avoid accidental (self-)pollinations. Pollen donors were kept separate from emasculated plants. Anthers with mature pollen from pollen donors were picked using forceps and carefully touched onto the stigma of the target receiving flower. This treatment was repeated every day until all receiving flowers (multiple per raceme) were pollinated between one to three times from the same donor plant with fresh pollen. Fruit development and seed set was assessed before the capsules split and shed their seeds. Modifications to this protocol are listed in the individual experimental details.

Experiment 1) Self-incompatibility. Previous UK reports show varying degrees of self-incompatibility for *H. non-scripta*, but typically conclude a low selfing rate to complete

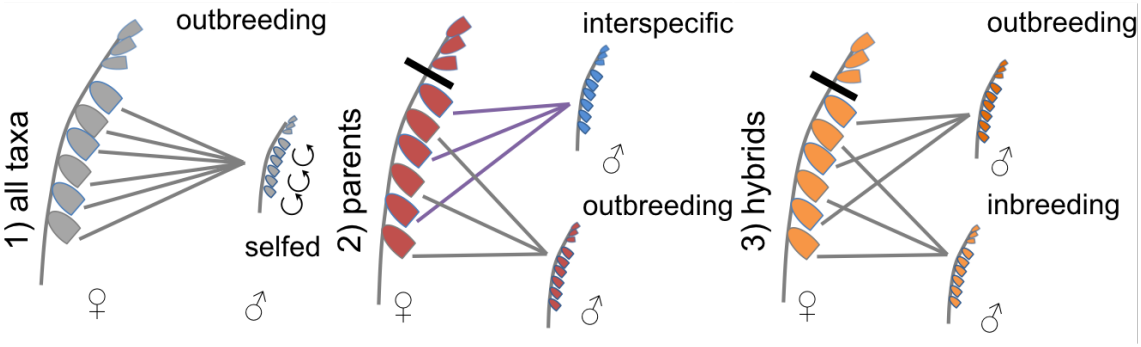


Figure 2.1 – Schematic of the three crossing experiments indicating which raceme (♀) received pollen from which pollen donor (♂). In experiment 1) the pollen donors were self-pollinated and their pollen also applied to several pollen receivers. In experiment 2) parent pollen receivers were outcrossed with inter-populational and inter-specific pollen. In experiment 3) hybrid pollen receivers were cross-pollinated from intra- and inter-populational pollen donors.

self-incompatibility (Corbet, 1998; Knight, 1964; Wilson, 1959b). The 'selfing' experiment aimed at investigating the selfing rate in *H. hispanica* and hybrids. Male pollen was applied on the stigmata of their own flowers (selfing) and on stigmata of several other pollen receiver from the same taxon (outcrossing) to confirm that the pollen was viable (Table 2.2). After the last hand pollination the racemes were individually bagged. About seven weeks after the crossings were completed, the swelling of the capsule as fruit initiation (yes = 1, no = 0) and the counts of fully developed seeds per capsule were documented.

Experiment 2) Hybrid Formation. The 'parent' experiment aimed at exploring the frequency of F1-hybrid formation by inter-specific crosses (Table 2.3). Further, it was determined if there is a directional difference in which species provides the pollen or ovule. Here, each receiving plant was pollinated by two different pollen donors, one originating from the other species (interspecific cross) and the other from the same species as receiving plant but from a different collecting site (outcrossing).

Experiment 3) Breeding system of hybrids. The 'hybrid' experiment aimed at testing the hypothesis that intra-population crosses (inbreeding) of hybrids show hybrid depression, while inter-population crosses (or outbreeding) of hybrids show hybrid vigour (Table 2.4). Each receiving plant was pollinated by two different pollen donors, one originating from a hybrid of the same collecting site (intra-population cross) and the other from a different collecting site (inter-population cross).

The parent and hybrid experiments were conducted simultaneously during three weeks from the end of April to early May in 2015. The position of treatment was altered within the raceme and the treatment was applied in pairs, meaning that always two flowers received pollen on the same day to minimise unequal resource allocation. Hand-pollination was repeated every second day until a flower started to wilt. By that, each flower received pollen two to four times. Further, the top flowers on stalks with more than 10 flowers were cut. Three weeks after the crossings were completed, the first capsules started to dry up. When a capsule began to turn brown and translucent, the number of un-developed ovules

or aborted seeds, and fully developed seeds per fruit were counted. This procedure allowed assessment of the rate of successful seed formation by a maximum number of ovules per capsule, which varies between flowers (25-30 ovules, Deroin, 2014). It was also recorded if a fruit began to swell subsequently to pollination. The number of carpels per capsule was also counted, which ranged from two to four, although three carpels are most common (Deroin, 2014).

Statistical analysis using mixed-effect model. The experiments outlined above were analysed using generalised linear mixed-effect models, which are a suitable model for non-normal data and allow to include biological 'nuisance' parameters in order to differentiate them from the parameter of interest, i.e. the effect of treatment.

The main interest of the experiments was to assess the differences in fruit and seed set due to differing treatments. A generalised linear mixed-effect model (GLMM) combines regression analysis for non-Gaussian 'response' data (therefore generalised linear model) with multiple 'explanatory' variables of either fixed (known) or random (unknown and require estimation) effects on the response variable (therefore mixed model). Random effects were modelled by adding intercepts to the model that assume the same trend of interactions with response variable but with differing starting points. Developing a GLMM requires three steps: 1) determine distribution for the response variable, 2) define the explanatory variables and their effects, 3) specify the link function between expected value of response variable and the explanatory effects for non-normal data (Nelder and Wedderburn, 1972). Possible explanatory variables are explained below.

For the counts of seeds per flower (i.e. the response variable) in experiment 1, a Poisson distribution was determined and a logarithmic link function was applied for the transformation (Bolker et al., 2009). For the presence or absence data if a fruit began to swell (i.e. the response variable), a binomial distribution with logistic link function was determined (Bolker et al., 2009).

For analysis of fully developed seeds and failed seeds per capsule (i.e. the response variable) in experiments 2 and 3 a binomial distribution was applied as it represents proportional data between successes and failures (Bolker et al., 2009). A logistic link function was applied for the transformation (Tables 2.3, 2.4).

Considered fixed and random effects based on bluebell biology and experimental design. The experimental setup and the biology of bluebells need to be considered to identify potential variables. The seed set per flower is biologically influenced by the position of the flower on the raceme (rank), the number of carpels per flower, and the fitness of the parents – especially the plant that is producing the seeds in these experiments (Corbet, 1998).

Bluebells are bulbous plants, which renew their bulb every season within the old one using up its resources. Therefore, they can be seen as a continuous, 'ageing' individual (Al-Modayan, 1993). The older a bulb, the more resources are stored as it accumulates resources over its life-span of the individual bluebell. The available resources per flowering season constrain the maximum number of flowers per raceme and counts of matured seeds. Since in the 2015 experiments each raceme received two treatments, their count data are

inter-dependent by the fitness of the receiving plant. In contrast, only a restricted number of plants was selected as pollen donors, but still their pollen fitness could influence the seed set. We cannot quantify the exact fitness of a plant, and hence included the parents of a cross ('mother', 'father') as random effects, if they contributed any variance.

The total number of ovules per flower is dependent on the number of carpels due to axile placentation and each carpel contributing 8-10 ovules to the placentary column (Deroin, 2014). However, some of the plants showed occasionally two or four carpels. Because the number of carpels biases the maximum count of seeds per flower, it was included as random 'nuisance' factor ('Ncarpels').

The flowers on a raceme open from bottom to top and become receptive sequentially. Therefore resource allocation also takes place sequentially and the number of produced seeds decreases with increasing rank of flowers, i.e. rank effect (Corbet, 1998). In the 2015 experiments, there might also be an interaction between rank and treatment of a flower, although the interaction was tried to be minimised by providing both treatments to the flowers on a raceme at the same time, and the order of the treatment was altered randomly between different receiving inflorescences. Consequently, rank was included as main fixed effect ('rank') in order to include the interaction term ('rank:treatment'). In the selfing experiment (2014), rank was included as a nuisance variable and therefore as random effect.

If more than one inflorescence per bulb was present, they could have differing resources affecting the seed outcome. The identifier of an inflorescence ('ID') was therefore included as random effect.

Selecting best GLMM and model validation. The parameter of the GLM models (GLMM) were approximated by maximum likelihood (Laplace Approximation) as used in the R package lme4 (Bates et al. 2014). To reach convergence of the model, the optimiser ('bobyqa') and run-time (100,000) needed adaptation from defaults. For example, a full GLMM that predicts the expected number of successful seeds as a function of all possible explanatory variables was defined in R using *glmer()* as follows:

$$
\begin{array}{lll}
\text{Logit(rate successes)} = & \text{flower(successes, failures)} \sim & \text{response variable} \\
& \text{rank + treatment +} & \text{fixed effects and} \\
& \text{rank:treatment +} & \text{their interactions} \\
& \text{(1|mother) + (1|father) +} & \text{random effects} \\
& \text{(1|Ncarpels) + (1|ID) +} & \\
& \text{(1|obs) + (1|fruit.ini)} & \text{correction overdispersion}
\end{array}
\qquad (2.1)
$$

Depending on the experiment, not all variables outlined contributed significantly to the expected outcome. Therefore, by removing each effect one-by-one, except treatment, a nested design testing the effect of each explanatory variable (and their interactions) was obtained. Note that, if an interaction between two variables is significant, both main variables should remain in the model even if they are not significant (Zuur et al., 2009). The analyses were performed separately for each experiment, and in addition the taxa were analysed individually for each experiment. Selection of the simplest model was

performed based on comparing residual deviances between the nested models using a chi-squared test (as well as AIC) for significant contribution of each effect (*drop1()* and *anova.glm()*; package stats; R Core Team, 2016). The best model was determined for each experiment, which is defined as the model with the smallest deviance, and the lowest number of explanatory variables that provide major variance (Tables 2.2 - 2.4). Lastly, model validation was visually assessed comparing the residuals of the models to the fitted data. A random intercept model was employed for each counted flower ('obs') if there was overdispersion in the data, defined here as the variance of the data being larger than its mean. In addition, the presence/absence of swelling fruit was included as random factor to deal with zero data. In modelling zero biased data, such as count data (Zuur et al., 2009), it improved model the fit by allowing a different intercept for failed flowers (i.e. zero seeds) compared to flowers that produced any other number of seeds.

The best model of each experiment was used to explore the effect of the treatment on the response variable. The marginal and conditional R^2 for GLMMs (Nakagawa and Schielzeth, 2013) were also reported to approximate how much variation in the observed data is explained by the fixed effects (R^2_{marg}) and by all variables (R^2_{cond}) in the chosen best model.

Germination of seeds from crossing experiments. Successful seed development alone is not evidencing embryo growth and seed viability. Seed viability was of interest in inter-specific crosses (F_1 hybrids) and hybrid crosses of later (natural) hybrid generations. According to Thompson and Cox (1978) the best germination rate of *H. non-scripta* can be obtained by a two-phase treatment: 1) pre-conditioning at high temperature (26 - 31°C) for four to ten weeks and 2) initiation of germination at 11°C. They observed that more than 80 % of the seeds germinated within three weeks after this treatment. Vandelook and van Assche (2008) obtained similar results in field experiments, where up to 86 % of *H. non-scripta* seeds have germinated after about three months of sowing. However, a cold stratification for bluebells has been suggested elsewhere (e.g. 4°C for ten weeks in Blackman and Rutter (1954), and 5°C for six weeks in Slade and Causton (1979); but see discussion in Vandelook and van Assche (2008) for the lack of seed dormancy in bluebells).

Seeds obtained in 2014 from the selfing experiment were pre-treated with 4°C cold stratification for 14 days to attempt breaking seed dormancy. Seeds were randomly selected to germinate from crosses with more than ten seeds in total. Potential contaminants like mould were removed by washing seeds in 70 % ethanol, then keeping them for 5 minutes in 5 % bleach. Subsequently, the seeds were rinsed with purified distilled water to remove residuals of bleach and ethanol. The seeds were spread on the damp filter papers in Petri dishes sealed with parafilm. Filter papers were also sterilised in diluted bleach and rinsed with purified water. All Petri dishes were placed in a 11°C cold room with a 12/12 hours light/dark regime. The experiment progress was examined monthly by counting the seeds with obviously prospering radicles. The germination experiment was terminated after a year. Finally, non-germinated seeds were cut in half to see if the seed was dead or an embryo still present (i.e. dormant).

Seeds obtained in 2015 from parent and hybrid experiments were not treated with a chilling phase because the germination rate in 2014 was very slow and below the reported

average germination success elsewhere. The sterilisation of the seeds was also not included in the protocol because it might have negatively influenced the germination rate in the previous year. Ten seeds per cross were randomly selected and placed onto damp, washed, filter papers into sealed Petri dishes. The Petri dishes were placed into a plant growth room at 22/19°C with a 12/12 hours light/dark regime respectively for 14 weeks as a warm conditioning phase following Thompson and Cox (1978). Afterwards the seeds were placed into a cold room at 11°C with also a 12/12 hours light/dark regime to initiate germination (start: 30th September 2015). Germination and growth progress was assessed at four time points: 20-12-2015, 20-01-2016, 18-04-2016, and lastly 30-08-2016. For each Petri dish the numbers of germinated, dormant, or dead seeds were counted. A seed was counted as germinated when the radicle emerged from the seed coat. A seed was counted as dead, if it was overgrown by mould and had become soft.

Table 2.2 – Summary of the samples and methods used in the selfing experiments, which tested the degree of self-incompatibility. Abbreviations: flws - flowers, fruit ini - fruit initiation, tot.poll.flws - total number of pollinated flowers, ns - *H. non-scripta*, hyS - hybrid South, and SD - standard deviation.

Experiment 1) Selfing	*H. hispanica*	hybrid North	hybrid South	*H. non-scripta*
Receivers2014	9	11	16	17
Tot. poll. flws	61	74	80	113
Selfed: plants(flws, min - max)	4(31, 5-13)	4(31, 4-9)	9(41, 2-8)	14(92, 3-20)
Outcrossed: plants(flws, min - max)	6:(30,3-8)	6(43,5-10)	7(39, 3-9)	4(21,4-6)
Replicated donors	none	6:	7:	4:
	483-C, 501-C applied to 2 ns, and 1 hyS: 3(15, 3-6)	401-B, 402-A, 403-D, 403-F, 495-B, 495-C	480-C, 482-A, 482-C, 492-B, 493-B, 494-C, 498-C	405C-11, 503-E, 503-G, 505-C
GLMM: presence/absence data of fruit ini per fruit (binomial distribution with logistic link-function)				
Random effects	mother, rank	mother, father, rank	mother, father, rank, flowers	mother, flowers
Fixed effects	treatment (crossed vs selfed)	treatment (crossed vs selfed)	treatment (crossed vs selfed)	treatment (crossed vs selfed)
GLMM: count data of seeds per fruit (poisson distribution with logarithmic link-function)				
Random effects	mother, rank, flowers	mother, rank, flowers	father, rank	mother, father, flowers
Fixed effects	treatment (crossed vs selfed)	treatment (crossed vs selfed)	treatment (crossed vs selfed)	treatment (crossed vs selfed)
Quantitative germination results				
Crosses that were germinated	8	8	8	8
Plants selfed (final success proportion ± SD)	2 (0.7 ± 0.28)	1 (0.8)	NA	2 (0.35 ± 0.071)
Plants outcrossed (final success proportion ± SD)	6 (0.82 ± 0.18)	7 (0.74 ± 0.32)	8 (0.84 ± 0.21)	6 (0.5 ± 0.37)

Table 2.3 – Summary of the samples and methods used in the parent experiments, which tested the frequency of hybrid formation. Abbreviations: flws - flowers, tot.poll.flws - total number of pollinated flowers, tot.poll.infl - total number of pollinated inflorescences, and SD - standard deviation.

Experiment 2) Hybrid formation	*H. hispanica*	*H. non-scripta*
Receivers2015	22	22
Total poll. infl	26	26
Tot. poll. flws	126	122
Intercrossed: plants(flws, min - max)	22(63, 2-7)	22(61,1-6)
Outcrossed: plants(flws, min - max)	22(63, 2-7)	22(61,1-6)
Donors2015: plants(poll.flws, min-max)	5(124, 8-57)	4(124, 15-47)
Plants used as pollen donor and receiver	490-C (9 flws were given its pollen; 4 flws received pollen)	395-A (15 flws were given its pollen; 6 flws received pollen)
GLMM: proportional data [0,1] of successful seeds per fruit binomial distribution with logistic link-function)		
Random effects	mother, father, flowers, inflorescence, fruit initiation	mother, father, flowers, inflorescence, fruit initiation, number carpels
Fixed effects	treatment (crossed vs hybrid), rank, and their interaction	treatment (crossed vs hybrid), rank, and their interaction
Quantitative germination results		
Crosses that were germinated	44 (of 52)	42 (of 52)
Plants intercrossed (final success proportion ± SD)	21 (0.96 ± 0.12)	21 (0.94 ± 0.12)
Plants outcrossed (final success proportion ± SD)	23 (0.93 ± 0.088)	21 (0.95 ± 0.12)

Table 2.4 – Summary of the samples and methods used in the hybrid experiments, which tested the breeding system of hybrids. Abbreviations: flws - flowers, tot.poll.flws - total number of pollinated flowers, tot.poll.infl - total number of pollinated inflorescences, and SD - standard deviation.

Experiment 3) Breeding system of hybrids	hybrid North	hybrid South
Receivers2015	16	10
Total poll. infl	21	11
Tot. poll. flws	112	58
Inbred: plants(flws, min - max)	16(59,1-6)	10(29,2-5)
Outcrossed: plants(flws, min - max)	16(53,1-6)	10(29,2-5)
Donors2015: plants(poll.flws, min-max)	6(107,6-33)	6(63,2-18)
Plants used as pollen donor and receiver	none	none
GLMM: proportional data [0,1] of successful seeds per fruit binomial distribution with logistic link-function)		
Random effects	mother, rank, inflorescence, fruit initiation	mother, rank, fruit initiation
Fixed effects	treatment (crossed vs selfed), rank	treatment (crossed vs selfed), rank
Quantitative germination results		
Crosses that were germinated	38 (of 42)	18 (of 22)
Plants inbred (final success proportion ± SD)	18 (0.97 ± 0.059)	9 (1.0 ± NA)
Plants outcrossed (final success proportion ± SD)	20 (0.98 ± 0.041)	9 (0.99 ± 0.033)

2.2.4 Ploidy assessment using flow cytometry and chromosome squashes

To explore the possibility of polyploid bluebell samples, flow cytometry (FCM) was used and supported by a few chromosome squashes. Leaves still attached to the bulb material from fieldwork were used for flow cytometry, which was performed at the Jodrell Laboratory of the Royal Botanic Gardens, Kew (UK). The nuclei counts were performed using a CyFlowSL Partec flow cytometer (Partec GmbH, Goettingen, Germany) fitted with a 100mW green laser (532 nm solid-state Cobolt Samba laser; Cobolt AB, Solna, Sweden). The nuclei extraction and staining method followed a two-steps protocol (Doležel et al., 2007): fresh material was co-chopped with the standard for reference and Tris $MgCl_2$ buffer in Petri dishes on ice. As reference standard *Allium cepa* 'Ailsa Craig' was used, which has a 2C value of 34.89 pg (Clark et al., 2016). RNase and propidium iodide (as stain) were added and the solution incubated for 10 minutes. Then the nuclei solution was filtered through a nylon mesh filter from tissue debris. The nuclei extract from leaf tissue was quite 'gooey' due to the high amount of polysaccharides in bluebell tissues (Brocklebank and Hendry, 1989; Simmonds, 2004). Consequently, the solution was diluted by four times buffer volume (i.e. 1 ml buffer to 0.25 ml solution). Possible interference of the secondary compounds (Price et al., 2000) was additionally tried to be kept to a minimum by storing the tubes on ice while the solution was running through the flow cytometer. There was only dried leaf material available for two collecting sites (BB-406, BB-488), which was first pre-soaked for 5 minutes in Cysteine Partec buffer and subsequently treated as described above. All resulting histograms of nuclei counts were manipulated using the FloMax 2.7 software (Partec GmbH, Goettingen, Germany) and the results were visualised and analysed in R version 3.3.1 (stats package; R Core Team, 2016).

To assess the ploidy level, at least 2500 nuclei were counted without replicates. Then the peak positions of the reference was compared with the bluebell samples, based on an expected peak for bluebells of 2C = 42.40 pg (Bennett, 1972). Between one to seven measurements were assessed to approximate the ploidy level of a collecting site (median = 3). In addition, if the sample's and standard's coefficient of variation (CV) was below five per cent, an individual genome size estimate was obtained as the ratio between the mean sample peak and the mean standard peak (from the histograms) times the standard's genome size. CV is a measure of how dispersed the nuclei counts in the histogram are around the mean peak; the smaller, the better. Of course, these estimates can still show a large uncertainty because of the low numbers of nuclei counts. Nevertheless, the data are suitable for ploidy level determination. For absolute genome size measurements at least two replicates of 5000 nuclei counts were performed. For *H. hispanica* fresh plant material was used from one sample growing in the Chelsea Physics Garden, London, that was originally sampled near Lisbon in 2008 (BB-188; Grundmann et al., 2010). Three replicates of the same nuclei suspension were measured. For an absolute genome size estimate of *H. non-scripta* from Spain, four individuals from the field collection (BB-505) were used. Here, two replicates for each individual were measured from the same nuclei suspension.

In order to support the ploidy estimates from FCM, chromosome squashes were performed for a few accessions. To prepare the chromosome squashes, actively growing root tips emerging from the bulbs in September 2013 were collected and pre-treated in 0.002M 8-hydroxyquinoline for 20 hours at 8°C. Subsequently, the solution was removed and re-

placed by the fixative (absolute ethanol + glacial acetic acid in ratio 3:1, respectively). The fixative was removed 48 hours later and replaced by 70 % ethanol for long-term storage at -20°C. This method followed Grundmann et al. (2010). To prepare the squashes, root tips were hydrolysed in 5M hydrochloric acid for 10 minutes at room temperature, and then dissected and macerated in 60 % acetic acid. The acid was removed with a tissue, and the DNA stained with a drop of filtrated 2 % certified aceto-orcein. Careful tapping of the glass lid squashed the cells and released the arrested chromosomes. Chromosome numbers were examined from metaphase preparations under a light microscope for eight different samples. Images of the karyogram were obtained at a 100x magnification using the software openlab (PerkinElmer) without size measurements.

2.3 Results

2.3.1 Spatial and ecological description of the hybrid zone area

In this study, a natural hybrid zone between the British bluebell, *H. non-scripta*, and *H. hispanica*, was sampled in the area between the west-southern foothills of the Cantabrian Mountains and the Duero basin at the border between Portugal and Spain. In total, 41 sites were visited and recorded (Figure 2.2), of which eight were already known from previous museum accessions (Table A.2). On average, 13 individuals were collected from 39 collecting sites, which include 141 *H. hispanica* individuals, 157 *H. non-scripta*, and 388 hybrid individuals. Living bulb material was also collected for almost all collecting sites (37 of 39; 131 bulbs) with a range of one to seven surviving plants for each site (Table A.2).

In general, specimens were found from altitudes of 534 to 1230 m either in deciduous *Quercus* and *Castanea sativa* forests with varying abundance of thorny shrubs such as *Cytisus, Rosa, Crataegus, Ulex*, and *Genista* in the understorey; or to a lesser extent on open rock sites and in river meadows. Bluebells seemed to avoid birch-dominated plant communities along streams and did not grow on sandy soils. The latter dominated in the East of the study area towards the Duero basin. Bluebells also seemed to prefer slopes but were rather tolerant in the context of vegetation cover but to a lesser extent to the density of grasses (which is likely caused by the avoidance of sandy patches). The British bluebell is known to avoid shade (especially when it inhabits forests) by flowering early before full foliage cover in the trees, and on open areas such as grasslands it can become outgrown by taller plants and competes over water (Blackman and Rutter, 1954; Ebuele et al., 2016).

The species identification in the field was based on morphological diagnostics of the inflorescence and the previously recorded occurrences ('old collection' locality information at the NHM). The transition between both parental taxa follows the Galician-Duero Mountains from North to South, which are bounded in the West by the Bierzo Basin, in the North by the Cantabrian Mountain ranges, and to the East and South by rather flat highlands of the Northern Meseta plain and the Duero Basin (Sobrino et al., 2007). The Galician-Duero Mountains – also called Galaico-Leonese Mountains, see Martín-González et al. (2012) – comprise several mountain peaks including Montes de León, Montes Aquilianos, Sierra del Teleno, Sierra de la Cabrera Baja, and Sierra de la Mina (Fig. 2 in Sobrino

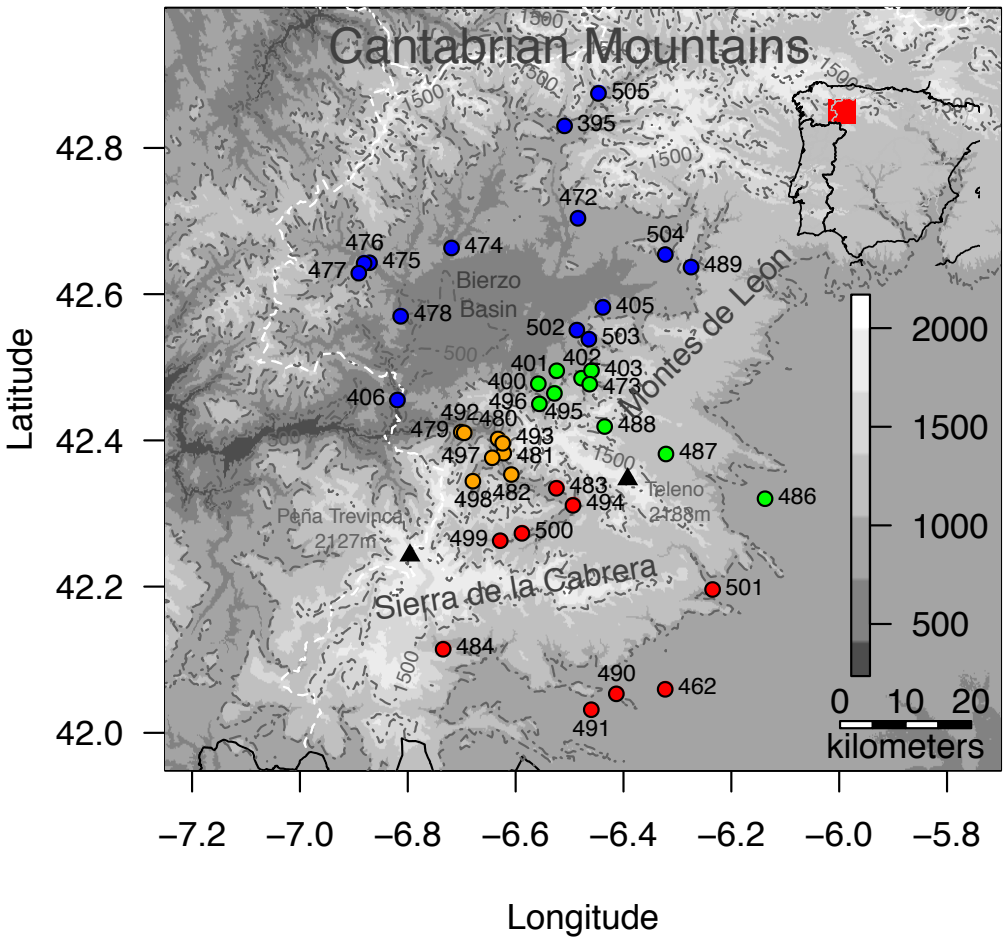


Figure 2.2 – Geological map of the study area (enlarged highlighted red region of Spain, top right) with 41 collecting sites visited during fieldwork. Each collecting site with its label is depicted as circle with the colours referring to the field identification based on morphology: blue – *H. non-scripta*, red – *H. hispanica*, green – northern hybrids, and orange – southern hybrids. The background is scaled to elevation profile as indicated by the legend in m. White dashed lines represent the border of Spanish local authority districts (West Galicia, East Castile and León) and the black line represents the Spanish border to Portugal.

et al., 2007), to name the ones relevant to this study area (Figure 2.2). Collecting sites of *H. hispanica* were found south of the hybrid zone and were separated by the Sierra de la Cabrera Baja with high elevational plateaus around 2000 m into northern – closer to hybrid sites – and south-eastern occurrences. The sites of *H. non-scripta* were sampled north of the hybrid zone centre between the southern foothills of the Cantabrian Mountains and the western foothills of the Montes de León in the East, and more distantly on the western side of the Bierzo Basin (Figure 2.2). For *H. non-scripta* the Bierzo Basin represented a distribution gap because nowadays most of it is transformed into agricultural and urban areas. However, this gap may be partly natural because fluvial sediments dominate this area, which are rich in ton particles and poor in gravel. Hybrid sites were found on either side of the ridge continuing from Montes Aquilianos to the Sierra del Teleno, which presents highest altitudes from 1850 m to 2185 m. Therefore, the hybrid zone centre runs from north-west to south-east over 90 km and is about 28 km wide in a north-south direction based on the presence of morphologically intermediate hybrid individuals. The southern hybrid sites were collected in the valley of the river Cabreira, which divides the Montes Aquilianos (N) from the Sierra de la Mina (S). At the eastern end of this valley *H. hispanica* sites were found. However, accessibility to collecting sites was limited by the lack of roads and infrastructure, which became increasingly scarce in high altitudes. In altitudes above 800 m – especially in the southern *H. hispanica* range – bluebells were more often at early flowering stages and therefore less conspicuous without the blue flowers (Table A.2).

The collecting sites usually had less than 100 individuals (52.5 %) but some had approximate numbers of up to 500 individuals (e.g. for *H. non-scripta* BB-502 to 503, or BB-475, 476, and 477), and it became evident that bluebells are potentially widespread if continuous habitat is available. Where bluebells grew in open forest understorey (e.g. BB-405, 475, 482, 483, 486, 504), they formed disjunct circular patches; in other areas they grew more widely dispersed. The density of plants is not as dense as to what is known for *H. non-scripta* in old-forest occurrences from the UK, for which they are an indicator species (Rose, 1999). Collecting sites very close to urban areas were not sampled to avoid cultivation forms. Very sturdy garden varieties of *H. hispanica* and the hybrids with broad leaves that are known from the UK (Wilson, 1956) were not observed.

2.3.2 Intermediate hybrid morphology

There were no sympatric parental species observed at any of the collecting sites. Instead, hybrid sites were identified based on individuals that expressed a range of intermediate morphologies between their parents. The collected hybrids were of similar size to their parents, all scented, and mostly coloured in varying shades of blue pigmentation with the exception of a few colourless plants. The colourless flowers likely carried mutations in anthocyanin pathway, as shown for selected bluebell cultivars that produce pink and white flowers (Stickland and Harrison, 1977). In natural British woods completely white plants are rare, occurring at densities of one plant per 1,000 – 10,000 (Riding, 1977; Wilson, 1959a). Without them being a major focus, white plants (i.e. no anthocyanin pigmentation in flowers and pollen) were discovered in two collecting sites, one of *H. non-scripta* (BB-475) and one of hybrids (BB-401). One white hybrid plant that was collected, BB-

401-A, successfully produced seeds and seedlings in the crossing experiment, indicating no reduced fecundity in the egg (Riding, 1977). In four hybrid sites, individuals with flowers were found that displayed either white (indicative of *H. non-scripta*), or blue (indicative of *H. hispanica*) anthers (e.g. BB-493 - Figure 2.3 E–G), or three anthers of each colour in separate whorls but displayed otherwise intermediate features, such as a bell-shaped perianth (BB-482, Figure 2.3 H).

In general, northern hybrids resembled mainly *H. non-scripta* with white to greenish pollen (Figure 2.3 N, O), bending tip of the inflorescence, and more strongly re-curved petal tips where the petals were mostly parallel and opened later at anthesis (Figure 2.3 B, J). In contrast, southern hybrids were more similar to *H. hispanica* with erect inflorescence (Figure 2.3 C, K), the perianth was campanulate, and the pollen green to dark blue (Figure 2.3 P, Q).

To quantify this field work observation, five morphological traits were scored for 160 individual plants from 39 different sites. The principal component (PC) analysis included only 119 plants with complete information (i.e. 74.4 %) resulting in 23 *H. hispanica* individuals, 38 *H. non-scripta* individuals, and 58 hybrids used for the analysis. Most of the variance in the morphological data was explained by the first component (PC1 = 79.0 %, PC2 = 9.8 %, PC3 = 4.6 %) and all morphological traits contributed equally to PC1. For PC2, habit was most decisive, and for PC3, it was the openness of the flower (Figure 2.5). The anther related characteristics (i.e. pollen colour, anther colour before dehiscence, pigmentation of anther epidermis) were strongly correlated (cor > 0.77), while habit presented lower correlations to all other variables (cor 0.59 – 0.69). The categorical classification of morphological traits was rather simplified and hence this analysis showed low resolution between individuals, e.g. the majority of *H. non-scripta* data points overlapped (Figure 2.4 top). Nonetheless, the analysis presented both parents widely spaced along the first PC with hybrids in intermediate positions (Figure 2.4 top). The variance of displayed morphology was larger for the hybrids than for their parental species (Figure 2.4 bottom). Interestingly, the centre of the PCA (Figure 2.4 top) was empty, indicating a lack of individuals with phenotypes representative of intermediate phenotypes, i.e. potential first generation hybrids. The transition of parental morphological traits followed latitude in clinal shape, with some hybrid samples overlapping along PC1 (Figure 2.4 middle). However, the means of PC1 for hybrids from the north and south were significantly different (Welch test: $t = -7.65$, $df = 54.41$, $p < 0.001$). Assuming the mountain peaks from Montes Aquilianos to Sierra del Teleno form a reproductive barrier, the overlapping hybrid phenotypes between the northern and southern samples could represent morphological plasticity of late generation hybrids, rather than homogenising gene flow (a hypothesis confirmed by genetics in chapter 4). Thus, the hybrid samples were split into two groups based on their occurrence on either side of the mountain range.

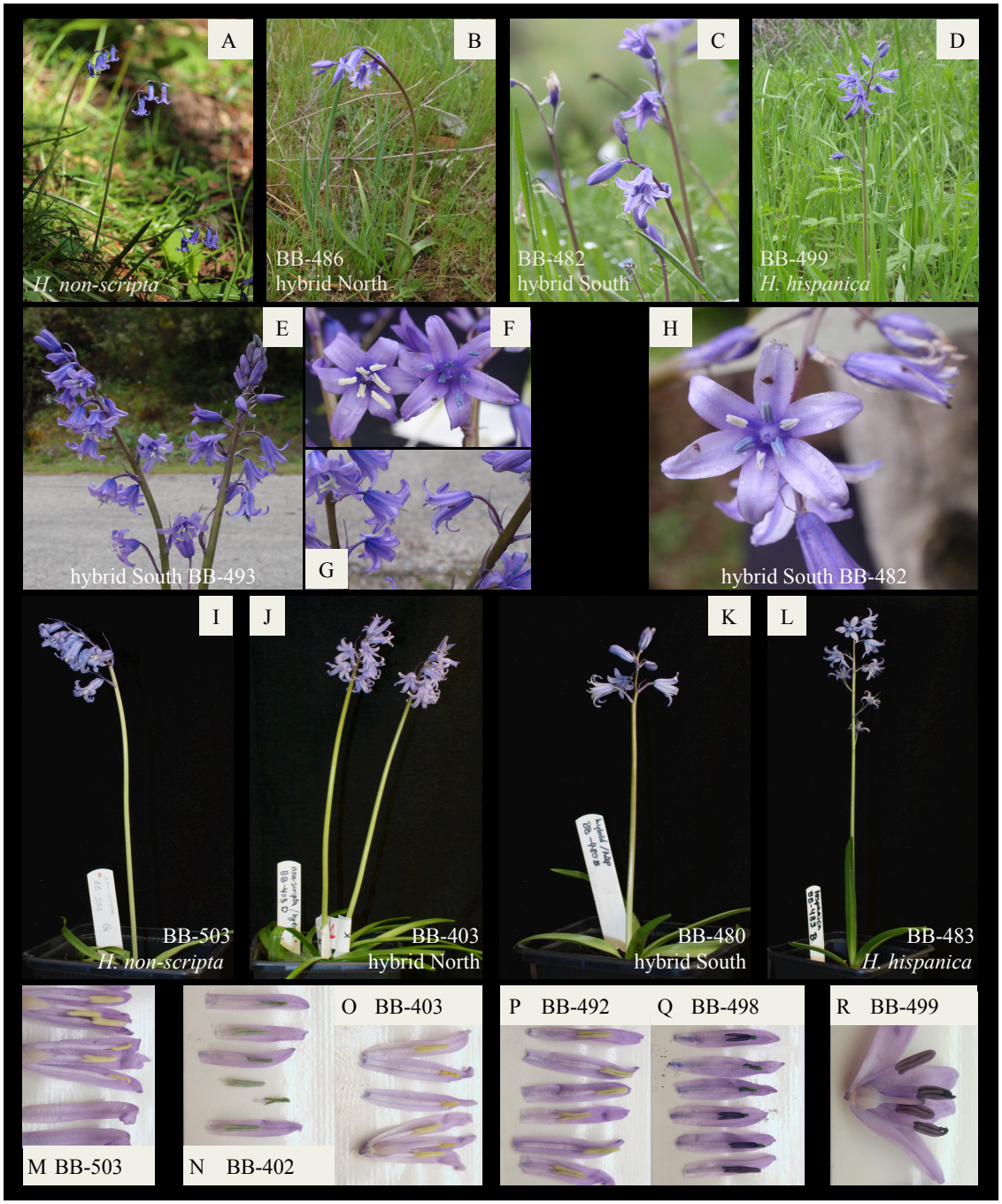


Figure 2.3 – Photographs of bluebells to illustrate the morphological variation: A – H were taken during fieldwork; I – R were taken in the laboratory.

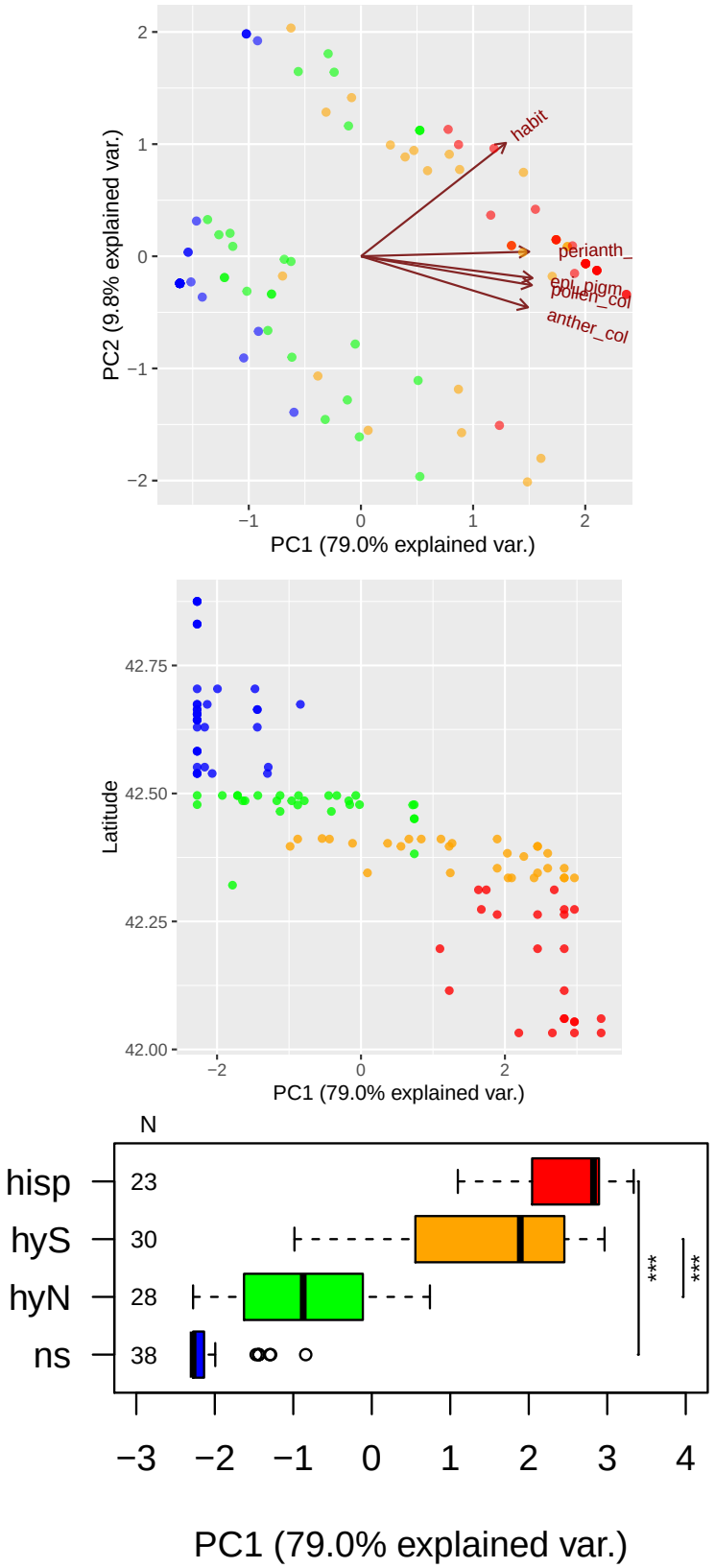


Figure 2.4 – Principal component (PC) analysis of five morphological characteristics for 119 individual plants, which included the taxa *H. non-scripta* – blue, *H. hispanica* – red, northern hybrids – green, and southern hybrids – orange. The top figure contrasts the first and second PC with arrows indicating the direction of contribution to the present variance in the data. In the middle figure, the morphological variation (by PC1) is plotted against latitude. The bottom figure illustrates the range of PC1 by the quantiles, median (thick lines) and outliers (circles) by taxon, as well as the number of N individuals per group, and the significance level of different means between the hybrid groups and the parental taxa (Welch test, *** = $p < 0.001$).

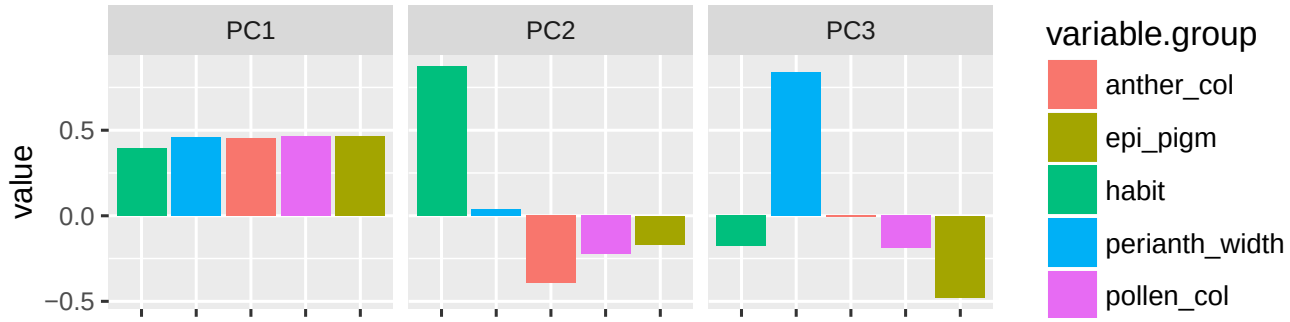


Figure 2.5 – Contribution of morphological characters (Eigenvalues) to variation in the first three principal components.

Table 2.5 – GLMM results for each hand-cross pollination experiment, indicating model parameter, i.e. degree of freedom of residuals – Df.resid and deviance. The conditional R^2 reports the percentage of variance explained by each model, and the marginal R^2 respectively for the effect of treatment. The effect of removing treatment from the model was tested using a χ^2 test (chisq).

Response variable	Exp.	N flws.	Model	Random effects	Fixed effects	Effect Treatment (Chisq)	P (> Chisq)	Df.resid	Deviance	AIC	R^2_{cond}	R^2_{marg}
Fruit initiation (selfing)	1)	328	binomial (logit)	mother, rank	8 treatments (taxa x cross)	73.446 (df=7)	p < 0.001	318	256.3	276.3	67.40%	46.30%
Seed count (selfing)	1)	328	poisson (log)	mother, rank, obs	8 treatments (taxa x cross)	118.58 (df=7)	p < 0.001	317	1284.4	1306.4	78.60%	54.10%
Seed proportion (parent)	2)	248	binomial (logit)	mother, father, Ncarpels, fruit.ini, ID, obs	4 treatments (taxa x cross)*rank	12.149 (df=6)	p = 0.0587	234	1210.5	1238.5	88.70%	0.139%
Seed proportion (hybrid)	3)	167	binomial (logit)	mother, rank, fruit.ini, ID	4 treatments (taxa x cross)	12.529 (df=3)	p = 0.0058	159	863.0	879.0	91.70%	0.323%

Table 2.6 – GLMMs were also performed for each taxon separately on the fruit initiation as response variable. As fixed effect only treatment was included for which the 'intercept' is indicative of the mean outcrossed effect in contrast to the 'estimate', which is indicative of the mean selfed effect in relation to the intercept. These values are given in logit space, and their standard error is given, which reports the accuracy of the effects.

Species / Treatment	Exp.	Samples	Fruit initiaion per fruit							
			Intercept	Estimate	SE	Df.resid	Deviance	z	P	R^2_{cond}	R^2_{marg}
H. hispanica											
Outcrossed vs selfed	1)	61	2.7 ± 1.4	-5.694	2.778	57	45	-2.05	0.04	77.1%	44.0%
Outcrossed vs hybrid	2)	126	NA								
H. non-scripta											
Outcrossed vs selfed	1)	113	1.7 ± 1.1	-3.17	1.291	109	116.3	-2.457	0.014	54.4%	16.3%
Outcrossed vs hybrid	2)	122	NA								
hybrid North											
Outcrossed vs selfed	1)	74	1.2 ± 0.7	-4.388	1.597	69	67.2	-2.748	0.006	60.4%	43.8%
Outcrossed vs inbred	3)	104	NA								
hybrid South											
Outcrossed vs selfed	1)	80	14.5 ± 6.1	-28.218	8.757	74	15.3	-3.222	0.001	99.8%	7.9%
Outcrossed vs inbred	3)	58	NA								

2.3.3 Experiment 1) Self-incompatibility

This experiment examined whether the hybrids and *H. hispanica* are self-incompatible as reported for *H. non-scripta*. To determine if selfing results in a failure of seed formation in bluebells, the following hand-cross pollinations were performed. A total of 17 receiving plants of *H. non-scripta* (113 pollinated flowers) were available, of which 14 plants were selfed (92 flowers) and the remaining outcrossed. Only four of the selfed donors were also used as donors in the outcrossing treatment. For all hybrids, there were 27 receiving plants (total of 162 pollinated flowers), of which 15 plants were selfed (total of 72 flowers) and the remaining outcrossed. Almost all selfed donor plants (14 of 15) were also used as pollen donors for the outcrosses. For *H. hispanica* there were only nine plants flowering. Due to the low number of *H. hispanica* individuals flowering simultaneously from distant collecting sites, the pollen from selfed donors could not be applied to other receiving *H. hispanica* individuals. Instead, to check that the pollen was viable, pollen used for selfing was applied to a range of other samples, including hybrid South and *H. non-scripta* (data not shown). The remaining six *H. hispanica* plants were used as receiving plants (30 flowers) and crossed with pollen from *H. non-scripta* and hybrid South. Therefore, the results from *H. hispanica* should be taken with caution based on the lack of replicated pollen donors and the overall limited number of crosses. See Table 2.2 for a summary of samples used for each taxon.

Overall, the treatment of outcrossing versus selfing provided a significant contribution to the presence or absence of fruit initiation, although the total model of the experiment explained only 67.4 % of the variance in the data (exp.1 in Table 2.5). Fruit initiation per flower and the count of seeds per flower were then analysed for each taxon separately using GLMM (Table 2.6 and 2.7). Random intercept models were applied for rank and for each mother and father plant, if they contributed any variance in the data. In addition, each flower was used as random effect to account for overdispersion (Table 2.2). The best GLM models accounted for different degrees of variance in the data with mostly

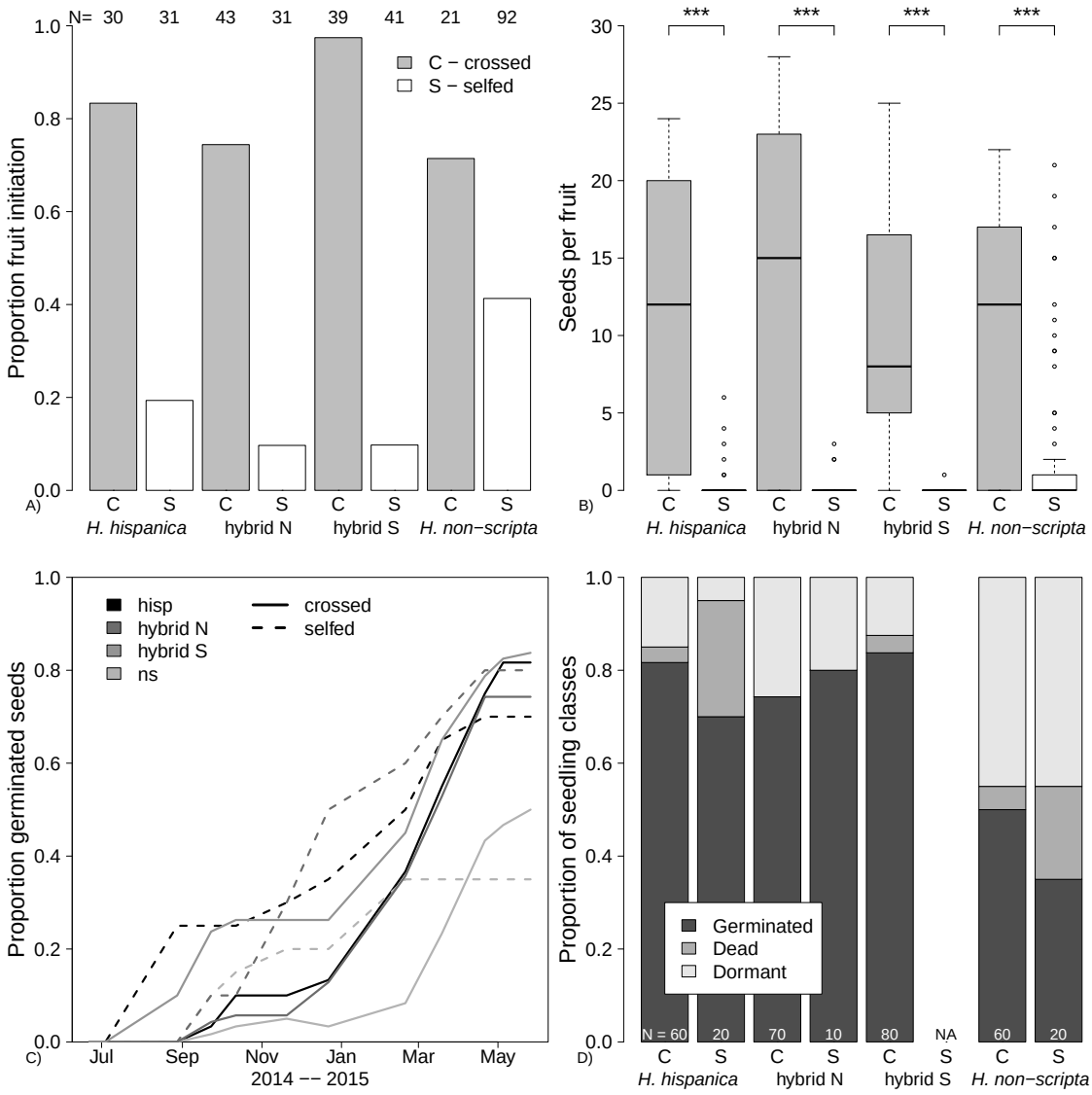


Figure 2.6 – Results of the selfing experiments: A) Proportion of initiated fruits per taxon and treatment with numbers of flowers on top (N). B) Range (quartiles and median) of seeds per fruit produced by treatment and taxon (raw data); significance level of GLMM output for treatment differences indicated on top (*** – $p < 0.001$). C) Time scale of germination by taxon and treatment. D) Final results of proportions of seeds that germinated, are dead or dormant with the total number (N) of analysed seeds.

Table 2.7 – GLMMs were also performed for each taxon separately on seed counts (exp. 1) or proportion of seeds per flower (exp. 2 and 3) as response variable. The 'intercept' is indicative of the mean outcrossed effect, while 'estimate' is given of the alternative treatment effect in relation to the intercept, although there are additional fixed effects (and interactions) included in the models (see Tables 2.2 – 2.4). The intercept and estimate values are given in logit space (except for exp. 1, which is in log space), and their standard error is given, which reports the accuracy of the effects.

Species / Treatment	Exp.	Samples	Seeds per fruit							
			Intercept	Estimate	SE	Df.resid	Deviance	z	P	R^2_{cond}	R^2_{marg}
H. hispanica											
Outcrossed vs selfed	1)	61	2.1 ± 0.5	-4.53	1.069	56	261	-4.241	<0.001	83.3%	54.9%
Outcrossed vs hybrid	2)	126	-5.5 ± 5.8	0.61	0.286	117	645.1	2.131	0.033	87.8%	0.134%
H. non-scripta											
Outcrossed vs selfed	1)	113	1.9 ± 0.8	-3.95	0.980	108	397.8	-4.036	<0.001	63.7%	26.3%
Outcrossed vs hybrid	2)	122	-3.5 ± 3.4	0.04	0.222	112	567.9	0.188	0.851	75.6%	0.0196%
hybrid North											
Outcrossed vs selfed	1)	74	1.3 ± 0.8	-5.54	1.252	69	343.6	-4.423	<0.001	83.3%	47.6%
Outbred vs inbred	3)	104	-5.7 ± 7.1	-0.31	0.098	98	551.3	-3.216	0.0013	91.0%	0.052%
hybrid South											
Outcrossed vs selfed	1)	80	1.9 ± 0.4	-6.06	1.013	76	244.6	-5.981	<0.001	87.7%	80.3%
Outbred vs inbred	3)	58	-1.2 ± 1.5	-0.17	0.132	53	273.6	-1.266	0.206	54.20%	0.076%

lower proportions reported by R^2_{cond} in the fruit initiation than in the seed set analyses (Table 2.6 and 2.7).

There is a significant difference between outcross-pollinated and self-pollinated treatment for each studied taxa in the proportion of flowers that initiated fruits (Table 2.6). It was highest for selfed plants of *H. non-scripta* (41.3 %), whilst it was amongst the lowest in outcrossed plants (*H. non-scripta* – 71.4 %) compared to the other taxa (Figure 2.6 A). For the selfed hybrids the fruit initiation was low (hybrid North – 9.68 %, hybrid South – 9.76 %) in contrast to the outcrossed hybrids (hybrid North – 74.4 %, hybrid South – 97.4 %) (Figure 2.6 A). Selfed *H. hispanica* individuals had a medium proportion of initiated fruits (19.4 %) in contrast to their outcrossed proportion (83.30 %).

Consequently, selfing resulted in significantly fewer seeds per fruit than outcrossing for all taxa (Figure 2.6 B, Table 2.7). The amount of variance explained by treatment ranged substantially from 26.3 % in *H. non-scripta* to 80.3 % in hybrid South. If seeds were produced by self-pollination, they occurred in some incidental plants of *H. hispanica* (one out of four plants with 2.8 seeds on average), and the hybrids (three out of 15 with 2.0 seeds on average). In contrast, for *H. non-scripta* seven out of 14 selfed plants produced an average of 5.4 seeds.

With regards to seed viability, ten random seeds from eight crosses were germinated for each taxon. These included only five crosses from selfing treatment because they generally produced less than ten seeds per cross. The germination rate was very slow (Figure 2.6 C), and even after one year, many seeds had not germinated - but they were not dead either (Figure 2.6 D). The proportion of germinated seeds was lowest in *H. non-scripta* (only 50 %). The experiment showed no strong difference in the proportions of germinated seeds that were produced in selfing or outcrosses. The slow rate of germination could be a representation of poor seed priming rather than low seed viability because the germination rate could be improved in the following year with a different pre-conditioning treatment.

To conclude, the first crossing experiment showed that some genetic self-incompatibility

is in place in *H. hispanica*, *H. non-scripta*, and their hybrids. In addition, the frequency of seed germination is not apparently reduced by hybrid or self pollination crosses.

2.3.4 Experiment 2) Hybrid formation

This experiment examined how frequently F_1 hybrids are formed and whether the proportion of seeds per flower is influenced by the parent that provides pollen or ovule. Fruit initiation was scored but not specifically analysed. The germination experiment was optimised to provide a long phase of warm pre-treatment to the seeds prior to germination, instead of cold temperature treatment.

A total of 26 flowering individuals of *H. hispanica* were used in crossing experiments, of which 22 plants were used as pollen receivers and five plants used as pollen donors; 490-C was used as both (Table 2.3). For *H. non-scripta*, there were 25 flowering individuals, of which 22 plants were used as pollen receiver and four plants used as pollen donors; 395-A was used as both (Table 2.3).

In total, 248 flowers were pollinated from 52 different inflorescences, which means that multiple inflorescences per plant were used (Table 2.3). Their seeds were individually collected by cross treatment per inflorescence and prepared for germination. Only one inflorescence did not set any fruit (i.e. two crosses) and another 17 crosses produced less than ten seeds (by treatment) and were therefore not germinated (Table 2.3). Consequently, a total of 85 different crosses were germinated.

A GLM model was established for the parent experiment, which included six different random effects, and treatment and rank, and their interaction as fixed effects (Table 2.5). Treatment consisted of four levels, namely inter-populational crosses for either species of *H. non-scripta* and *H. hispanica*, and inter-specific crosses for each species as pollen receiver. This model explained 88.70 % of the variance in the data, but fixed effects explained only 0.14 % of the variance in the data. Overall, the interaction between treatment and rank was significant (deltaAIC = +4.9, $X^2_{df=3} = 10.91$, p = 0.012), but not just treatment itself. The GLMs were repeated for each taxon (Tables 2.3 and 2.7). Only in *H. hispanica* did the comparison between outcrossed versus inter-specific treatment show significant positive effect on the proportion of seeds (p = 0.03; Table 2.7). Contrarily, the raw data of proportional seeds per flower did not show a strong difference between treatments for *H. hispanica* (mean outcross and inter-specific were both 0.56), neither for *H. non-scripta* (mean outcross 0.57 versus inter-specific 0.54). Based on the median of the fitted data *H. hispanica* produced slightly more seeds in inter-specific crosses (median fitted outcross 0.65 versus inter-specific 0.62), as observed for *H. non-scripta* (median fitted outcross 0.55 versus inter-specific 0.58) but without significant difference (Figure 2.7 A). The fitted values of the model showed that overall *H. non-scripta* produced fewer seeds than *H. hispanica* (53.2 % vs. 55.95 %, respectively, Figure 2.7 A). Therefore, there is a very low (post-zygotic) reproductive barrier between both species. Based on the overall lower number of produced seeds per flower for *H. non-scripta*, this might indicate that there is a benefit for hybrid formation when *H. hispanica* provides the ovule. However, the results were very close and the pairwise comparisons of treatments between taxa in the full model of the experiment showed no significant difference (data not shown).

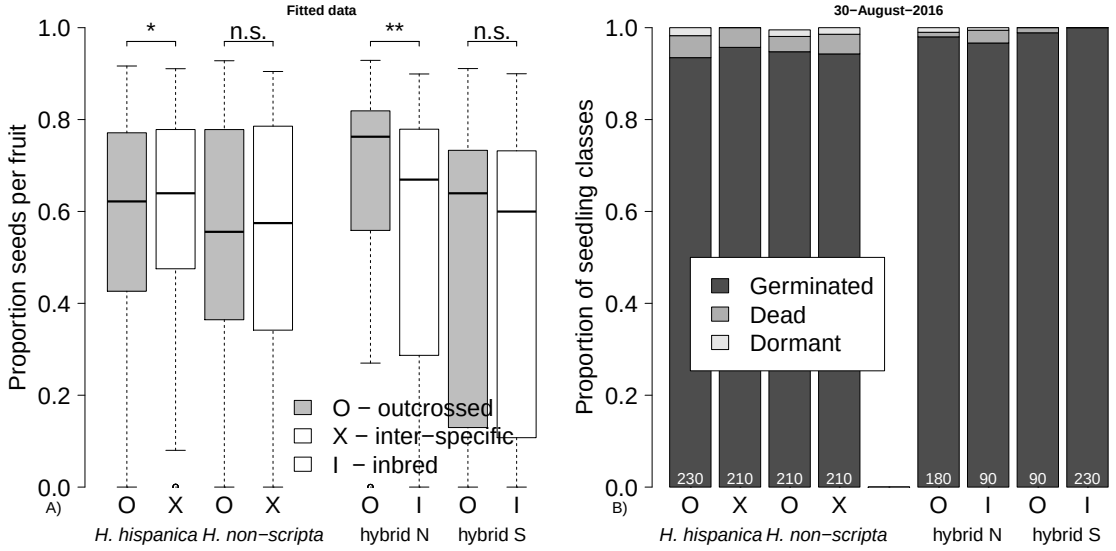


Figure 2.7 – Results of the parental and hybrid crossing experiments: A) Represents the range of fitted values of the proportion of matured seeds per fruit (quartiles, median and outliers as circles) from the GLMM for each treatment and taxon. Significance of treatment differences by taxon are indicated at the top (* – $p < 0.05$, ** – $p < 0.01$, n.s. – not significant). B) Final proportion of germinated, dormant or dead seeds by treatment and taxon; the total number of seeds analysed is shown at the bottom.

Within two weeks from transferring the Petri dishes into the 11°C cold room the seeds started to germinate. However, the seedlings were first counted only three months after the onset of germination, and already then 93.6 % of all seeds had germinated. By the end of the germination experiment (11 months later) the proportion had increased to 94.5 % germinated seedlings with only 1.16 % dormant and 4.34 % dead seeds (Figure 2.7 B). Consequently, there is high seed viability for both parental species and their reciprocal F_1 hybrids. In addition, 94 % of all seedlings formed their first bulb and 20.81 % showed new growth at the end of the experiment. The seedlings are expected to survive longer in their Petri dishes.

2.3.5 Experiment 3) Breeding system of hybrids

This experiment aimed to determine how well hybrids produce seeds and whether the geographic distance to pollen donor mattered. The breeding system (outbreeding versus inbreeding) of natural hybrids between *H. non-scripta* and *H. hispanica* was tested. Inbreeding here means pollination by a sample from the same collecting site as the receiving plant (within ca. 10 m distance). Outbreeding, in contrast, is referred to by pollination with a sample from a different collecting site with a minimum distance of 1 km (max. 20 km).

There were a total of 38 hybrid plants flowering in 2015, of which 22 belonged to the northern hybrid region and 16 to the southern hybrid region (Table 2.4). As in the previous experiments, multiple inflorescences were present for only a few plants. For hybrid North, 16 plants were pollinated with the pollen from six different donors. The treatment was applied to 112 different flowers. For hybrid South, ten plants were pollinated by also six donors, which resulted in only 58 different flowers (Table 2.4).

The seeds per inflorescence and treatment were individually collected and prepared for germination. Eight crosses failed to produce sufficient seeds to germinate at least ten seeds (Table 2.4). Consequently, a total of 56 different crosses were germinated (Table 2.4).

A GLM model was established (167 observations of individual flowers; df: 159, deviance: 863) that included treatment as fixed effect and four random effects (SI 2). Treatment consisted of four levels, namely outbreeding crosses for either northern or southern hybrid group, and their inbreeding crosses. The developed model explained 91.40 % of the variance in the data, but the fixed effect (i.e. treatment) only explained 0.33 % of the variance in the data. Overall, treatment had a significant effect (deltaAIC = +6.53, $X^2_{df=3}$=12.53, $p < 0.001$). For hybrid North there was a strong negative difference between outbreeding versus inbreeding ($p < 0.01$). For hybrid South there was also a negative effect between outbreeding versus inbreeding but not significant ($p = 0.216$). The fitted values of the model showed that hybrid South produced fewer seeds on average (46.74 %) than hybrid North (59.8 %) and also presented a larger variation (Figure 2.7 A).

Seedlings were first counted three months after the onset of germination and 97.68 % of all seeds had already germinated. By the end of the germination experiment (11 months later) the proportion of germinated seedlings increased to 97.86 % with 1.08 % dormant, and 1.06 % dead seeds (Figure 2.7 B). In addition, 96.61 % of all seedlings formed their first bulb and 30.9 % showed signs of new growth from the bulb. Thus, there is high seed viability for both inbred and outcrossed hybrids. The germination occurred more efficiently for the hybrid crosses (98.4 %) than for the parental crosses (94.6 %). The improvements to seed priming were very efficient at increasing the germination rate from 67.8 % in 2014 to 96.5 % in 2015, suggesting that the lower germination rate in 2014 was due to poor priming.

To conclude, inbreeding resulted in fewer seeds (54.5 % and 45.4 %, respectively) than outbreeding (65.1 % and 48.1 %, respectively) in hybrid North and South. Inbreeding was here applied by crossing samples from the same collecting site. Therefore, hybrids, especially hybrid North, could experience inbreeding depression as a result of genetic load. More importantly, this experiment showed that hybrids have a high fecundity and seed viability comparable to the levels of their parental taxa under laboratory conditions.

2.3.6 Large genome and homoploid hybrids

A total of 133 runs were performed using flow cytometry and only one individual (BB-406-A) failed to be adequately estimated for genome size, probably because it was from dried material. The outcome of dried tissue was generally poor. Bennett and Smith (1976) published a genome size estimate of 2C = 42.40 pg (2n, Fe method) for the British bluebell, *H. non-scripta*, under its synonym *Endymion non-scriptus*. Therefore, the expected 2C value of diploid bluebells would peak at a 1.2-fold larger position than the 2C peak position of the standard reference (*Allium cepa*, 2C = 34.89 pg) in the nuclei counts histogram (Figure 2.9 A). The mean peak position of a triploid nuclei count would be 1.5-fold larger than a diploid bluebell in the histogram. From runs with a good coefficient of variation (CI $\leq$ 5 %, i.e. 70 of 133 runs = 52.63 %) the ratio between the sample and standard mean peaks was estimated to obtain a distribution of good estimates (Figure 2.9 B, mean $r_{(Sample/Standard)} = 1.38$). All successful runs but six fell within this range of the diploid

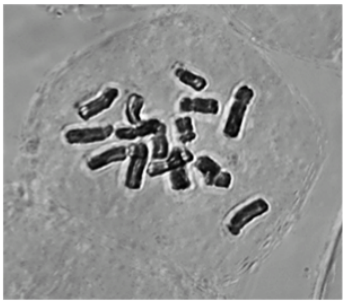


Figure 2.8 – Diploid *H. non-scripta* from Spain (BB-505-D) with 2n=16 chromosomes viewed at 1000x magnification.

distribution. Five of the six outliers had poor measurements, which can explain the shift to extremely large ratios. The sixth outlier was one replicate of the sample BB-188-CPG, which showed a good coefficient of variation. Because the other two replicates of BB-188-CPG had similarly high results ($r_{(Sample/Standard)} = 1.42$), this could represent a real result (see below).

As a result, diploidy was inferred for 114 individuals of field material with fresh material, six individuals that provided only dried material, and eight individuals that were sampled fresh from the Chelsea Physics Garden. Accordingly, the sampled 38 different collecting sites in the hybrid zone can be assumed to present diploid specimen. Multiplying the ratio with the genome size of *Allium cepa* (2C = 34.89 pg; Clark et al., 2016) resulted in a genome size estimate per sample (Table A.1). The chromosome squashes of all eight samples (four *H. non-scripta*, two *H. hispanica*, two hybrids) showed no irregularities as 16 chromosomes were counted for several root tip cells in metaphases of mitosis. Consequently, a homoploid hybrid zone was assumed.

2.3.7 Genome size estimates for *H. non-scripta* and *H. hispanica*

The absolute genome size for *H. hispanica* (BB-188-CPG) was 2C = 49.63 ± 0.2 pg (one sample, three replicates). This sample was also cytogenetically analysed by Grundmann et al. (2010) presenting 16 chromosomes. The absolute genome size estimate of *H. non-scripta* from Spain (BB-505) was 2C = 47.44 ± 0.21 pg (four samples, two replicates each). Its chromosome squashes also presented 16 chromosomes (Figure 2.8). The absolute genome size estimates between *H. hispanica* (BB-188-CPG) and *H. non-scripta* (BB-505) were significantly different (Two Sample t-test: $t = 15.84$, $df = 3.82$, $p < 0.001$).

In order to compare these parental genome sizes to the remaining individuals from the hybrid zone, the mean genome size of all runs with good measurements (i.e. CI ≤ 5 %) was estimated by taxon. The absolute genome size estimate of BB-188 was significantly larger than expected from the mean genome size of all other supposedly pure *H. hispanica* sampled in northern Spain, i.e. 2C = 48.29 ± 0.42 pg (12 runs from 6 different sites; Two Sample t-test: $t = -5.3$, $df = 13$, $p < 0.001$). The absolute genome size of BB-505 was significantly smaller than the mean genome size of all supposedly pure *H. non-scripta* collected in Spain, i.e. 2C = 48.24 ± 0.48 pg (19 samples from 8 different collecting sites, CV ≤ 5 %; Two Sample t-test: 4.56, $df = 25$, $p < 0.001$). In contrast, there was no significant difference between the mean estimates of the other samples of *H. non-scripta* and *H. hispanica*

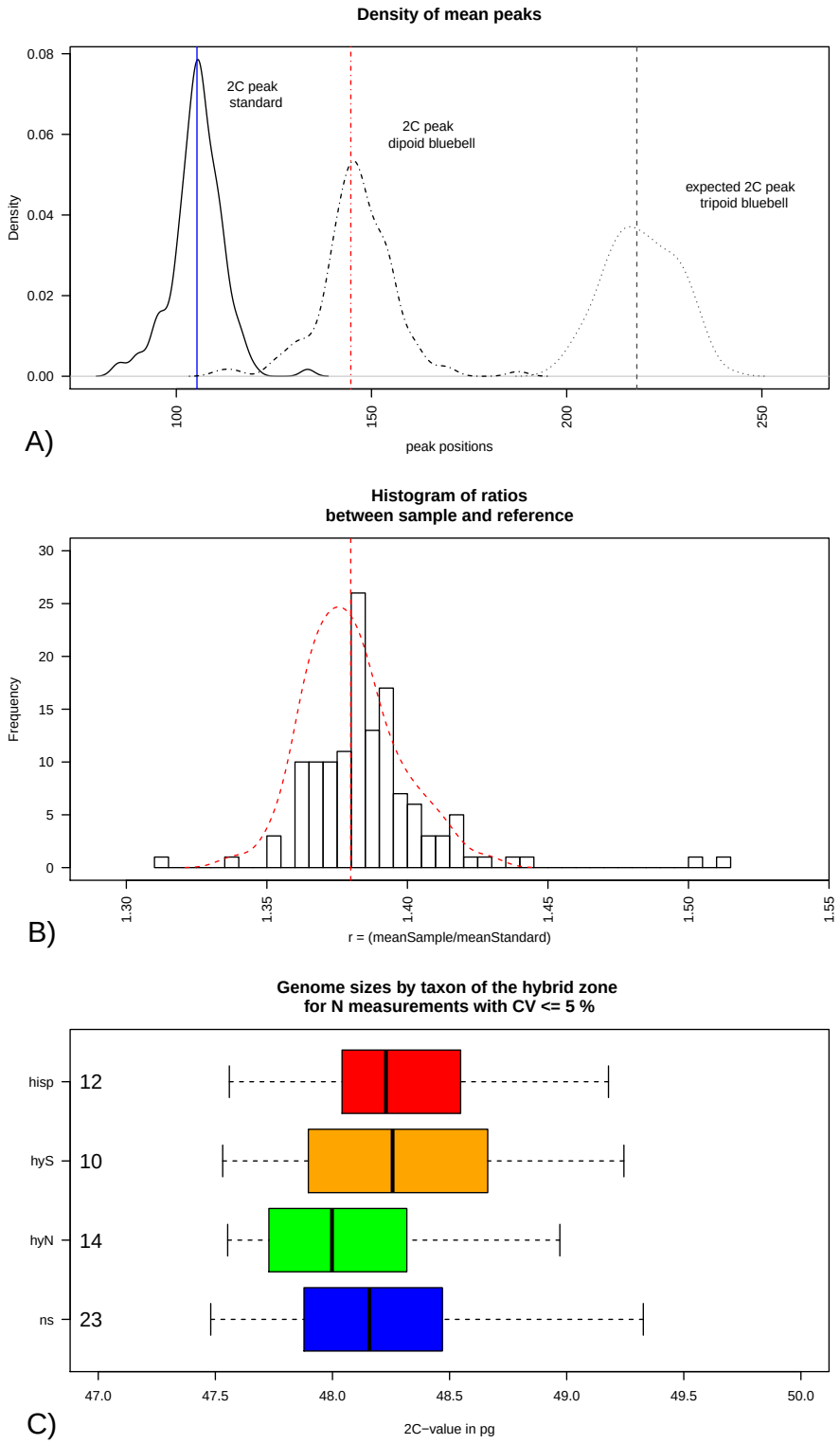


Figure 2.9 – Results of ploidy and genome size assessment. A) Overview of obtained mean peak positions from the FCM for *Allium cepa* (reference standard, blue, CV $\leq$ 5 %), diploid bluebells (CV $\leq$ 5 %), and the expected peak positions for triploid bluebells (grey); and their means (vertical lines). B) Histogram of ratios between bluebell and reference mean peaks and in red the distribution of good measurements and its mean (same values as used in A) to identify outlier measurements for ploidy determination. C) Range of good genome size estimates for each taxon, vertical lines represent quartiles and median (0 %, 25 %, median, 75 %, and 100 %) and the number on the left are the amount of measurements.

from northern Spain (Two Sample t-test: t = 0.32, df = 25.94, p = 0.75). For comparison, the mean genome sizes for northern hybrid individuals, i.e. 2C = 48.03 ± 0.42 pg (13 samples from 6 sites, CV ≤ 5 %) and southern hybrid individuals, i.e. 2C = 48.13 ± 0.69 pg (11 samples from 4 sites, CV ≤ 5 %) fall within the variance of the parental measurements (Figure 2.9 C). Based on the low count of nuclei these results were considered as fairly inaccurate. However, they represent for the parental taxa as well as for the hybrids similar genome sizes between 47.6 – 48.6 pg (95 % CI). In contrast, the absolute genome sizes for *H. hispanica* and *H. non-scripta* represent extreme values along local genome size estimates.

2.4 Discussion

2.4.1 Redefining the hybrid zone area

Hybrids between *H. non-scripta* and *H. hispanica* have been discovered in Northern Spain along the Cantabrian Mountains from chloroplast sequences and allozyme data (Steve W. Ansell pers. comm.) and in the central Sierras of Spain, where the discrimination between both taxa seems to be problematic (see discussion in Grundmann et al., 2010). In the present study, a thorough data sampling along the Galician-Duero Mountains (south of the Cantabrian Mountains) added 33 new parental and hybrid sites and re-defined the location and extent of a natural bluebell hybrid zone. Living samples provided crucial data addressing variation in morphology and genome sizes of samples along the hybrid zone. Performing hand-cross pollinations 1) an incomplete but strong self-incompatibility for *H. hispanica*, *H. non-scripta* and their hybrids was shown, 2) a high inter-crossability between *H. non-scripta* and *H. hispanica* with a potential of heterozygote advantage with *H. hispanica* as mother was shown based on seed development and germination rate, and lastly 3) high seed set and seedling viability with a potential for outbreeding advantage was shown for their natural hybrids. *H. non-scripta* is well studied in its British distribution range in respect to life history (Blackman and Rutter, 1954; Thompson and Cox, 1978; Wilson, 1959a), ecological requirements (Merryweather and Fitter, 1995; Sims et al., 2014; Vandelook and van Assche, 2008), and even its chemical compounds for commercial use (Kato et al., 1999; Simmonds, 2004; Thoss et al., 2012). Contrarily, *H. hispanica* was not granted such intense research. Therefore, for simplifications in the discussion, information about *H. non-scripta* is collectively applied as bluebell biology, assuming both closely related taxa share main biological traits including life history and ecological preferences.

2.4.2 Importance of the Galician-Duero Mountains in post-glacial colonisations

Northwest Iberia is a large and geographically complex area, where two different biogeographical regions, the Eurosiberian and Mediterranean, are considered to be in contact along the southern foothills of the Cantabrian Mountain range extending further west along the Galician-Minho Mountains to the mouth of the Duero river into the Atlantic Ocean (Sobrino et al., 2007). The highly variable relief and climatic patterns of northwestern Iberia led to diverse ecosystems (Ramil-Rego et al., 1998), which are dominated

by broad-leaved forest with various oak species (Buide et al., 1998; Díaz-Maroto and Vila-Lameiro, 2007). The Galician-Duero Mountains are characterised by a montane climate in higher altitudes that get more annual rainfall with decreasing temperature, surrounded in lower altitudes by a meso-Mediterranean bio-climate (the Bierzo and Duero Basins) with drier summers (Ramil-Rego et al., 1998). In this study, bluebells were mostly found in deciduous *Quercus* forests between 500 – 1200 m altitude, which probably belong to the *Genisto-falcatae-Quercetum pyrenaicae* phytosociological association found in the Bierzo Basin, and associations with chestnut such as *Linario triornithophorae-Quercetum pyrenaicae* present in the Galician eastern mountain and *Linario triornithophorae-Quercetum petreae* present in the Cantabrian Mountains (Ramil-Rego et al., 1998). Further, bluebells are sensitive to summer droughts and soil frost in winter based on their germination requirements (Thompson and Cox, 1978) and bulb investigations (Merryweather and Fitter, 1995), and show a preference for acidic soils in the British Isles (Blackman and Rutter, 1954) and in northern Spain. These climatic requirements fit the rather humid and drought sensitive communities described above (Ramil-Rego et al., 1998).

Both parental species, *H. hispanica* and *H. non-scripta*, were found to be closely related in chloroplast sequences and it was further suggested that the younger diversification of the genus *Hyacinthoides* might be a consequence of the Pleistocene glaciation cycles (Grundmann et al., 2010). Quaternary oscillations throughout the Pleistocene and Holocene over the last 2.6 my have been shown to strongly affect the distribution of Europe's fauna and flora (Gómez and Lunt, 2006; Hewitt, 1999, 2011; Schmitt, 2007). A popular example of Quaternary consequences in the Iberian Peninsula is the allopatric speciation and the survival of one lineage in a southern refuge, the grasshopper subspecies *Corthippus parrallelus erythropus*, which came into contact due to post-glacial expansion (after the last glacial maximum) with its conspecific, *Corthippus parrallelus parrallelus*, in the Pyrenees (Bella et al., 2007). Examples of plant species originating in the Iberian Peninsula have been shown elsewhere (Comes and Kadereit, 2003; Gómez and Lunt, 2006). Since the last glacial maximum (20.5 ka) the Northwest of the Iberian Peninsula has been glacier free and has experienced a few cycles of colder periods concordant with global climate oscillations in the Holocene (Sobrino et al., 2007), although over larger time scale the Holocene has provided rather stable conditions. Consequently, the species assemblages (based on pollen data) have changed a few times on local ranges resulting in turnovers of vegetation types at given altitudes with slight differences between sites of the Cantabrian Mountains within their biogeographic region (Sobrino et al., 2007). Since the end of the cold Younger Dryas (13.7 ka) oak and mixed deciduous forest species have re-colonised the Cantabrian Mountains and progressively the interior Galicio-Duero Mountains from the coastal areas in the North but mostly from the South-West (Iriarte-Chiapusso et al., 2016). Under the assumption of possible Pleistocene allopatric speciation between *H. non-scripta* and *H. hispanica* and their survival in Northwest Iberian refuges, the bluebell hybrid zone described would represent a secondary contact (Grundmann et al., 2010) after the Youngest Dryas. The time frame for the Younger Dryas in North-western Iberia was determined to have occurred about 13.7 ka, based on species assemblages in soil cores (Sobrino et al., 2007). Specifically both species might be expanding into the Galician-Duero Mountains, which provide a complex territory with suitable habitat.

2.4.3 Hybrid formation and fitness

Hybrids between *H. non-scripta* and *H. hispanica* are well known in the UK and were even given their own nothospecies epithet, *H. x massartiana* (Geerinck, 1996). Given the ease of hybridisation between bluebell cultivars and *H. non-scripta* in its northern distribution range, the high inter-crossability across the hybrid zone might not be surprising. But this study is closing a critical research gap of quantitative evidence in bluebell conservation management (Kohn et al., 2009) because, so far, no crossing experiments have been published that measured how frequent F_1 hybrids are produced. Although hand-cross pollinations ignore pre-zygotic isolation mechanisms, they still provide valuable information. Crossing experiments found application to study breeding systems of invasive plants with strong nature conservation impact (e.g. Ward et al., 2012).

The common expectation of inter-specific hybridisation was that hybrids present reduced fitness compared to their parents (Arnold and Martin, 2010; Barton and Hewitt, 1985), which can be due to genomic incompatibilities between the genomic backgrounds of the parental species (Orr, 1996), ecological maladaptation, or reduced mating success (Lindtke et al., 2012). In contrast, hybrid advantage can occur due to new genomic combinations that enhance adaptation to environments intermediate or distinctively different from the parents, as shown in spruce (De La Torre et al., 2015). Higher hybrid fitness can also occur in early generations of hybrids due to elevated heterozygosity by which depleted genetic diversity such as after a bottleneck event could be overcome (Kirk et al., 2005). The slight increase of seed set in F1-crosses and the high hybrid fecundity shown here for bluebells is therefore unusual for hybrid zone studies from a traditional perspective of hybrid zone models. However, hand-pollinations in bluebells have been shown to be much more efficient in bluebells than open pollination (Corbet, 1998). Consequently, it is questionable how much the small treatment effect differences in the parental crosses are realised as hybrid advantage in the natural environment – especially when considering floral traits as potential pre-zygotic barriers to gene flow (as discussed below). Further studies that include backcrosses and second hybrid generation crosses could allow to test the hypothesis of hybrid breakdown after potential heterosis effects in later generations and their isolation from the parental species.

The outcrosses of northern hybrid individuals showed a significantly higher seed set and heterosis might play a role in this bluebell hybrid zone. But the low differences between outbreeding and inbreeding treatment might also present sufficient gene flow between and within collecting sites so that the distances between sites were not sufficient to capture population differentiation.

Self-pollinated flowers mostly produced no or low seed counts per flower (Corbet, 1998) implying self-incompatibility. Avoiding selfing leads to a higher genetic differentiation within populations, which could be beneficial in bluebells to maintain genetic diversity given that they also exhibit vegetative reproduction by bulblets (Wilson, 1959a). The vegetative reproduction enables geophytes to form large colonies and secure propagation (Decocq and Hermy, 2003) and their higher relatedness (in very local space) might also support the establishment of self-incompatibility to maintain genetic diversity.

2.4.4 Morphologically intermediate bluebell hybrids with implications for reproductive barriers

It is a challenge to compare the morphology of the hybrids in the natural hybrid zone in Spain with those hybrids recorded elsewhere. This is because bluebells have been quite popular as ornamental plants in Western Europe (Bleeker et al., 2007; Kohn et al., 2009; Page, 1987; Quené-Boterenbrood, 1984) and the delineation between the true *H. hispanica* (originated in the Iberian Peninsula) and cultivars summarised under the '*Spanish bluebell*' in regions where no natural *H. hispanica* occurs, led to a confusion of morphological characteristics for both, *H. hispanica* and hybrids (Booy et al., 2015; Geerinck, 1996; Grundmann et al., 2010; Kohn et al., 2009; Page, 1987). However, in several regions along the distribution range of *H. non-scripta* morphologically intermediate hybrid populations were recognized based on vegetation census or herbarium material: Belgium (Geerinck, 1996), the Netherlands (Ietswaart et al., 1983; Quené-Boterenbrood, 1984), and the British Isles (e.g. Dines, 2005; Stickland and Harrison, 1977). What these reports have in common to our natural hybrid zone is the frequent observation of hybrids and their intermediate morphology to the parental taxa. In contrast, the *H. hispanica* individuals collected in Northern Spain were scented (as were the hybrids and *H. non-scripta*), which was disputed for the '*Spanish bluebell*' cultivar (Geerinck, 1996). The symmetrical transition of morphological characters along latitude with segregation into two hybrid groups is concordant with the direction of gene flow across the hybrid zone based on the parental distributions and the Sierra del Teleno ridge as potential barrier to gene flow. The morphological diversity in hybrids is a consequence of segregation and recombination between parental genomes (Lowe and Abbott, 2015), which seems to present low linkage given the array of shapes and colours in bluebell hybrid flowers.

Gene flow between collecting sites of both hybrid groups is most likely maintained by pollen dispersal because bluebell seeds show no adaptations to dispersal (Knight, 1964) and migration rates for *H. non-scripta* in a transplant experiment in Belgium were very low with 0.6 – 6 cm/y (van der Veken et al., 2007). The shapes of flowers, for plants that rely on pollen dispersal via pollinators, are often a key in divergent selection and form reproductive barriers (Charlesworth and Charlesworth, 2000). For instance, phenotypic divergence between two ecotypes of *Mimulus aurantiacus* in California is driven by selection on floral trait by pollinator preferences, while the genomic divergence is very low, probably due to the early phase of the speciation process (Stankowski et al., 2016). Now, for bluebells we demonstrated that there is a low post-zygotic barrier (explored by intercrossability and seed germination), but the different flower shapes might pose a pre-zygotic barrier (Campbell et al., 2002). The pollinator groups named for *H. non-scripta* include bumblebees (*Bombus*), long-tongued hoverflies (Syrphidae, e.g. *Melanostoma*) and solitary bees (e.g. *Anthophora*) (Blackman and Rutter, 1954; Kohn et al., 2009; Willmer, 2011). The open perianth of *H. hispanica* and the bell-shaped hybrid perianth would allow to be visited by larger and smaller, or long- and short-tongued pollinators, while the narrow-tubular flowers of *H. non-scripta* attract smaller and long-tongued pollinators. Consequently, pollinators that reach for *H. non-scripta* can also pollinate hybrids and *H. hispanica*, while gene flow is restricted in the reverse direction. Yet, successful

inter-specific pollination depends on the range of pollen dispersal by its pollinators and contemporaneous flowering of the conspecifics (e.g. Marques et al., 2007; Michalski and Durka, 2015). In the field, we noticed that the flowers were in earlier stages at higher altitude and further south for *H. hispanica*, and northern sites such as BB-472, and BB-505 for *H. non-scripta* – without critical assessment. During the crossing experiments samples from southern Spain (BB-491, 492, and 462) were flowering later, so that inter-specific crosses were a timely issue within two weeks. It was further reported that *H. hispanica* flowers later than *H. non-scripta* for the Netherlands (Knight, 1964) and for Belgium (Geerinck, 1996). In the British Isles *H. hispanica* appears to show a strong gradient in flowering time from south to north and perhaps also from low to high altitudes (pers. comm. Fred Rumsey, NHM London). However, primary data from Northern Spain that tested differences in flowering time between the parental taxa are absent. This variation in flowering time might therefore play a role in sympatric parental populations (Kohn et al., 2009; Michalski and Durka, 2015), especially in the British Isles, but possibly less so in the Spanish hybrid zone where sympatric parental populations were absent.

Questioning the risk of hybridisation between the native *H. non-scripta* and 'alien' taxa (generalising *H. hispanica* and hybrids), Kohn et al. (2009) argued that alien bluebell taxa within 1-2 km distance to native British bluebell populations would provide considerable genetic interactions. In contrast, several studies showed that long distance pollen dispersal by insects is possible over 1800 m (e.g. Millar et al., 2014), although dispersal range often shows low mean distances, e.g. 17.2 m within population for *Campanula thyrsoides* (Scheepens et al., 2012). These two examples describe the commonly shown leptokurtic distribution of pollen dispersal, where most pollen is dispersed (very) close to the source and rapidly declines with increasing distance (Ellstrand, 1992). However, pollen dispersal is dependent on the insect behaviour (and therefore the insect group) and its plant interaction. Pollen dispersal would need to cross about 100 km to facilitate direct gene flow between *H. non-scripta* and *H. hispanica* plants (and the area is strongly structured by the several mountain peaks).

2.4.5 Homoploid hybrid zone with large variation of genome size estimates

The cytogenetics work showed that all samples examined were diploid, which leads to the conclusion of a homoploid hybrid zone for bluebells in this region. Contrarily, in the UK triploids (2n = 24 chromosomes) were found amongst *H. non-scripta* and *H. hispanica* (supplement of Grundmann et al., 2010). These triploids represent larger size and more vigorous plants and might have been selected as horticultural plants (Grundmann et al., 2010; Wilson, 1958). Such vigorous plants were not observed in the study area. At the same time, we cannot exclude the hypothesis that homoploid hybrids also occur in the UK (work is in progress at RBGE).

The absolute genome size estimate of *H. hispanica* from southern Portugal (2C = 49.63 ± 0.2 pg) and *H. non-scripta* from northern Spain (2C = 47.44 ± 0.21 pg) present extreme ends of the spectrum of all measured samples. The previous record of genome size for *H. non-scripta* obtained by Bennett and Smith (1976) was much smaller (2C = 42.4 pg)

but it was also obtained using a different method (Feugen densitometry; see Baack et al. (2005) for discussion) and probably the sample is from the UK (Bennett, 1972). Here, both new absolute estimates should present reliable results because the FCM approach included replicated measurements of the same nuclei solution with high nuclei counts, and standard and sample were co-chopped as is commonly done (e.g. Clark et al., 2016). In addition, the samples BB-505-D and BB-188 (maintained at the CPG) are confirmed diploids based on chromosome counts (this study; Grundmann et al., 2010). For the remaining samples of the hybrid zone (for which the focus was primarily on determining ploidy level) the genome sizes vary considerably between different collecting sites, but sample variation is similar between taxa. The large variation in genome estimates could be due to inhibitors (such as secondary compounds) acting on propidium iodide (Price et al., 2000), and due to the lower numbers of nuclei counts, especially in contrast to the absolute genome sizes obtained for *H. non-scripta* and *H. hispanica*. Besides, the genome size of a species is generally expected to be more or less constant, if there is enough gene flow between individuals of a population (Greilhuber, 1998). But intra-specific genome sizes can vary to up to 10 % (Baack et al., 2005), and interspecific hybridisation may increase genome size and its variation in hybrids and backcrosses (Vega et al., 2013). For instance, in a three-way hybridisation between loosely related species of orchids (genus *Epidendrum*), F2 hybrid generations showed the strongest increase of genome size, while the parents had similarly small genomes (2C-value from 3.72 to 3.98 pg; Vega et al., 2013). Such strong difference between genome size of hybrids and their parents was not observed in this study, although the parents present significantly different genome sizes. Rather, the genome size of bluebell hybrids was intermediate. Homoploid hybrid zones between the sunflower species *Helianthus annuus* (2C = 7.23 pg) and *H. petiolaris ssp. fallax* (2C = 6.68 pg) also exhibited no increase of DNA content, although there was a maternal effect (Baack et al., 2005). In the case of sunflowers the hybrid fertility is very low (< 1 %; Rieseberg, 2000), probably due to genomic re-arrangements, and therefore the hybrids undergo a strong genetic bottleneck (Baack et al., 2005). Regarding bluebells, given the high inter-specific crossability in the natural hybrid zone, strong genomic incompatibilities are not likely and the observed intermediate genome size of the hybrids not unusual.

Lastly, confirmation of the diploidy of samples will be convenient in designing genetic markers and analysing single nucleotide polymorphisms, which is more challenging in polyploid samples (Dufresne et al., 2014).

2.5 Outlook

To conclude, in this study a natural hybrid zone in the Galician-Duero Mountains was defined by thorough sampling and morphological characterisation of the transition between *H. non-scripta* and *H. hispanica*. Cross-pollinations provided essential information about the breeding system and the ease of inter-specific hybridisation, which led to a number of hypotheses that can be tested with genomic studies using the collected DNA material:

1. Using nuclear and organelle genetic markers, how much gene flow can be observed between both parental species? Are there barriers to gene flow?

2. Can loci be determined that are involved in reproductive isolation between both parental species, e.g. by targeting candidate genes for flowering traits and cyto-nuclear interactions?

3. Can genetic evidence be found for heterosis-driven introgression?

2.6 Acknowledgements

Thanks to the team that was involved with the field work: Prof Harald Schneider, Prof Andrew Leitch, Prof Richard Nichols, and Alexandre Blanckaert. Further, I am grateful to Jasmin Zohren, Sarah Seco-Leite, and Paul Fletcher for their help and care taken of the plants and seedlings during my absences.

Chapter 3

Developing a multi-gene SNP marker set to study hybridisation in plants with large sized genomes

3.1 Introduction

Genomic resources for large genome species. Species with genome sizes of 1C > 14 Gb (1C is the amount of DNA in haploid unreplicated nucleus) are defined by Leitch et al. (2005) as being 'large' genomes. Genomes of this size are problematic for next-generation sequencing (NGS) technologies because of the huge volumes of DNA sequence data that must be obtained, and because of large numbers of repeats these genomes harbour; repeats which severely restrict the quality of assemblies. However, gene numbers often vary little between species, for example the conifer *Picea abies* (28,354 genes, 1C = 19.6 Gb) has a similar number of genes compared to the angiosperm *Arabidopsis thaliana* (25,498 genes, 1C = 0.15 Gb), despite a > 10-fold difference in genome size (Michael and Jackson, 2013).

Currently complete, published genomes are largely restricted to model plant species and well-studied crop species, all with genome sizes smaller than 1C = 5 Gb (Feuillet et al., 2011; Michael and Jackson, 2013). When assembly projects have targeted species with large genomes, for example the conifer *Picea abies* (1C = 19.6 Gb) and wheat *Triticum aestivum* (1C = 17 Gb), assemblies were restricted to about 61 % and 22 % of their genomes respectively (Michael and Jackson, 2013). Besides the sequencing costs, the assembly problems mean that species with large genome size species are underrepresented in whole genome sequencing studies (Kelly and Leitch, 2011). In consequence, they are also underrepresented in population genetics studies that apply NGS technologies (Egan et al., 2012; Schoebel et al., 2013). This is a problem, particularly because species with large genomes are expected to be more frequently threatened by extinction (Vinogradov, 2003) and by the effects of climate change (Knight et al., 2005) than species with small genomes. Thus, many species with large genomes are of conservation interest. Consequently, there is a need to develop effective strategies to obtain population genetics markers from genomic DNA samples, the function of this paper.

Genomic marker for the scale of population studies. Here, a marker set for genic regions of two closely related species with large genomes was developed, with the aim to provide diagnostic alleles that can distinguish them and characterise multiple hybrid generations in varying environmental backgrounds. Hybridisation between species is a common evolutionary phenomenon with hybrids occurring in 40 % of plant families (Whitney et al., 2010). In order to trace hybridisation and the transfer of genes between species through hybridisation and backcrossing (i.e. introgression) the marker set needs to be scalable to population genetic studies, where many individuals can be sampled simultaneously (Twyford and Ennos, 2011). The challenges in developing molecular marker for population genetic studies have been summarised by Schlötterer (2004), but briefly, markers should follow Mendelian inheritance patterns, be reproducible, scorable across all individuals with low numbers of null-alleles, and maximise the amount of information for minimum costs and efforts. NGS technologies promises to facilitate cost-effective access to molecular markers for populations (Soltis et al., 2013).

Several strategies are available, but they are all problematic in one way or another. RAD sequencing is a method that uses enzymes to randomly chunk up DNA and has become very popular for population genomic studies (Davey and Blaxter, 2010; Davey et al., 2011). Unfortunately, the risk of null alleles increases with the size of the genome because of the number of potential digestion sites and increased demands of sequencing coverage (Arnold et al., 2013). In addition, the enrichment for genic regions becomes less efficient with genome size and, critically for species with large genomes, the costs remain prohibitively high and certainly cannot be achieved with a modest budget ($<$ £15,000).

Transcriptome-based approaches also appear, at first sight, an obvious way to follow populations of species with large genomes, since it requires no prior sequence knowledge and can also be used to characterise genes. Indeed RNAseq is one of the most widely used genome reduction strategy (Cronn et al., 2012). But for population genetic studies, which might involve fieldwork, the collection of mRNA is only possible from fresh material, and so it excludes access to herbarium specimens. Material from the field can be stored using preservatives/anti-degradants for RNA (such as RNAlater, Ambion Inc., Austin, TX), but the composition of the transcriptome itself (i.e. the proportion of transcribed genes) varies with time of day, season, tissue and life cycle stage of the collected material. Usually the first step, after the transcriptome is obtained, would be *de novo* assembly, which without a reference genome – as is expected for non-model organisms with large genomes – becomes a challenge due to families of gene isoforms (i.e. all transcript that form alternative splicing variants of the same gene, Reddy et al., 2013). Furthermore, the huge volumes of data that are obtained from population studies of sequenced transcriptomes become a challenge to store and handle.

Alternatively, the transcriptome itself can be used to search for polymorphic markers, such as simple sequence repeats and single nucleotide polymorphisms (SNP) (Guo et al., 2015; Salgado et al., 2014; Vatanparast et al., 2016). Searching for SNPs has advantages over other genetic changes because there is the potential to compare polymorphisms under selection with those alleles fluctuating in a population by genetic drift (Helyar et al., 2011; Seeb et al., 2011). Single SNPs selected by certain marker properties can be targeted by amplicon design using traditional PCR methods. In combination with multiplexing

several loci in a sequencing library, such PCR-based enrichment becomes cost-effective (Cronn et al., 2012). Amplicon sequencing from PCR is well established, it promises a high target specificity and uniformity in sequencing, can be applied across different sample batches, is reproducible and once the primers are verified and stocked, the technology becomes relatively cost-effective for increased number of samples, because NGS costs are low (Cronn et al., 2012).

Especially, micro-fluid reactor technologies have been developed to independently amplify few products and avoid challenges such as off-target priming, primer-dimer formation, and varying quantities of individual amplicon concentrations (Cronn et al., 2012). For instance, the Fluidigm Access Array System provides a 48 x 48 array, which enables 48 genomic samples to be amplified against 48 different PCR primer combinations. Due to the possibility to simultaneously amplify 10 different target sequences in each reaction, there is a theoretical potential maximum of 480 different target sequences per sample to amplify. However, when pooling the samples to a single lane for Illumina sequencing, there is a limit of 384 unique barcodes possible (i.e. pooling eight different 48.48 Access Arrays).

For the re-sequencing of the designed target regions, high sequencing depth is important to reliably identify the polymorphic markers and to perform genotyping.

The primers themselves also need to be designed carefully for effective multiplexing, limiting targets to genomic regions of low complexity and low mutation rates (Cronn et al., 2012). However, using a transcriptome reference promises success by taking advantage of conserved flanking regions for primer development (Vendramin et al., 2007).

Aim of this study. This study is designed to find multiple SNP markers in reference transcriptomes that can then be identified in genomic DNA from multiple individuals using multiplexing PCR and NGS technology. Genetic markers were developed from two species of European bluebells in the genus *Hyacinthoides* Heist. ex Fabr., namely the British bluebell *H. non-scripta* and its closely related species *H. hispanica*. These two species are diploids with large genome sizes (1C = 23.2 and 24.3 Gb, respectively; chapter 2). Their distribution ranges overlap naturally in central sierras of Northern Spain, where they form hybrids (Grundmann et al., 2010). In contrast, an introduction of the '*Spanish bluebell*' to the British Isles led to on-going hybridisation between both species – a process suggested to change the natural diversity of the British bluebell irreversibly (Kohn et al., 2009). Thus, the aim was to establish genetic markers suitable to test the hypothesis of adaptive introgression through inter-specific hybridisation in a natural hybrid zone besides the study of anthropologically-enforced hybridisation in the British Isles.

3.2 Material and Methods

3.2.1 Plant material

For RNA-seq, four different libraries from three European bluebell species were used: *Hyacinthoides hispanica* (BB-356, BM000864264; BB-339, BM000864247), *Hyacinthoides non-scripta* (BB-411, BM000864320), and *H. paivae* (BB-130, BM000865089). These plants were collected in the Iberian Peninsula (2005 and 2008, see Table A.2 for more details) and grown at the Chelsea Physics Garden in London, UK (CPG) until RNA

was extracted from premature inflorescences still enclosed by the bulb. The developed markers were tested on a few samples of *H. non-scripta*, and *H. hispanica* across their species range as a proof of concept that the marker development worked. Hence, for amplicon re-sequencing, 75 individuals of *H. hispanica* (36 individuals) and *H. non-scripta* (39 individuals) collected from 21 different collecting sites in Portugal, Spain, France and the UK were used (Figure 3.1). These were mostly obtained from field collections in 2013/14 (see chapter 2), or from silica material of museum accessions that have been collected between 2005 and 2009. In 14 cases, fresh DNA material was collected from living plants in the Chelsea Physics Garden London to provide a broader sampling (Tables A.1 and A.2). However, because these plants were initially collected in 2008, and were grown without careful curation, the sampled individual might not have been the original material. To test those samples' identity, one replicate of living (leaf material harvested 2013) and silica material (collected 2008, BB-393-3), and mixed resources for different individuals from collecting site BB-397 were included.

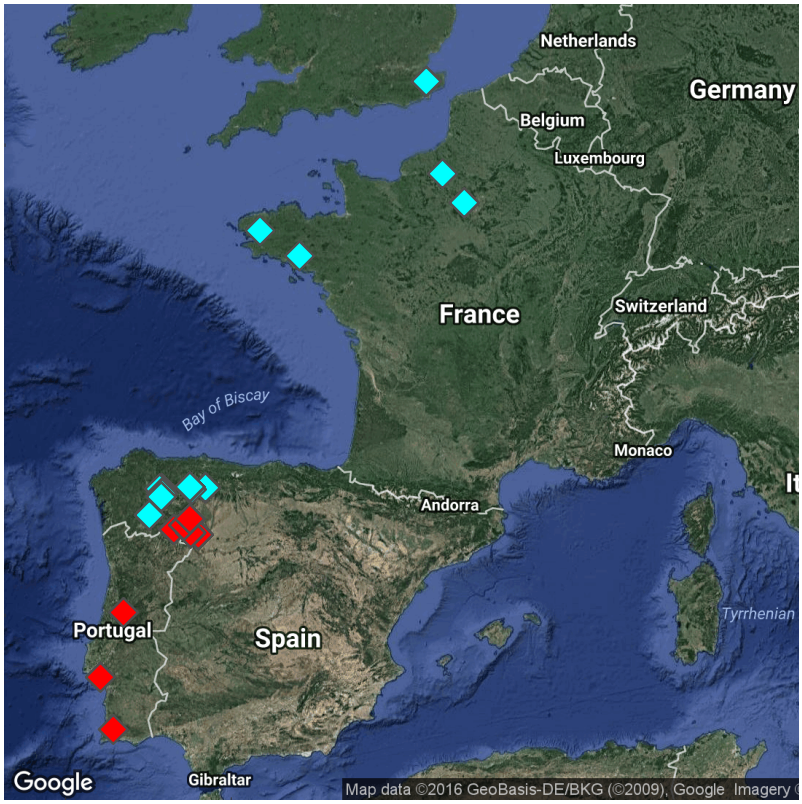


Figure 3.1 – The study includes 75 samples collected across Europe. For *H. non-scripta* 39 individuals were gathered from the UK, France and Spain (blue), and 36 *H. hispanica* individuals were gathered from Spain and Portugal (red).

3.2.2 Pipeline overview

In brief, the workflow spanned (1) pre-processing of RNAseq data; (2) *de novo* assembly of the transcriptome for each bluebell species; (3) comparison of the three transcriptomes and to the proteome of a closely related monocot, in order to identify and annotate shared (orthologous) genes; (4) variant discovery; (5) selection of target SNPs within 300 target

sequences, mostly from nuclear genes, but also from mitochondrion and plastid genes, for which primers were designed; and lastly (6) amplicon re-sequencing of those sequences from genomic DNA (Figure 3.2).

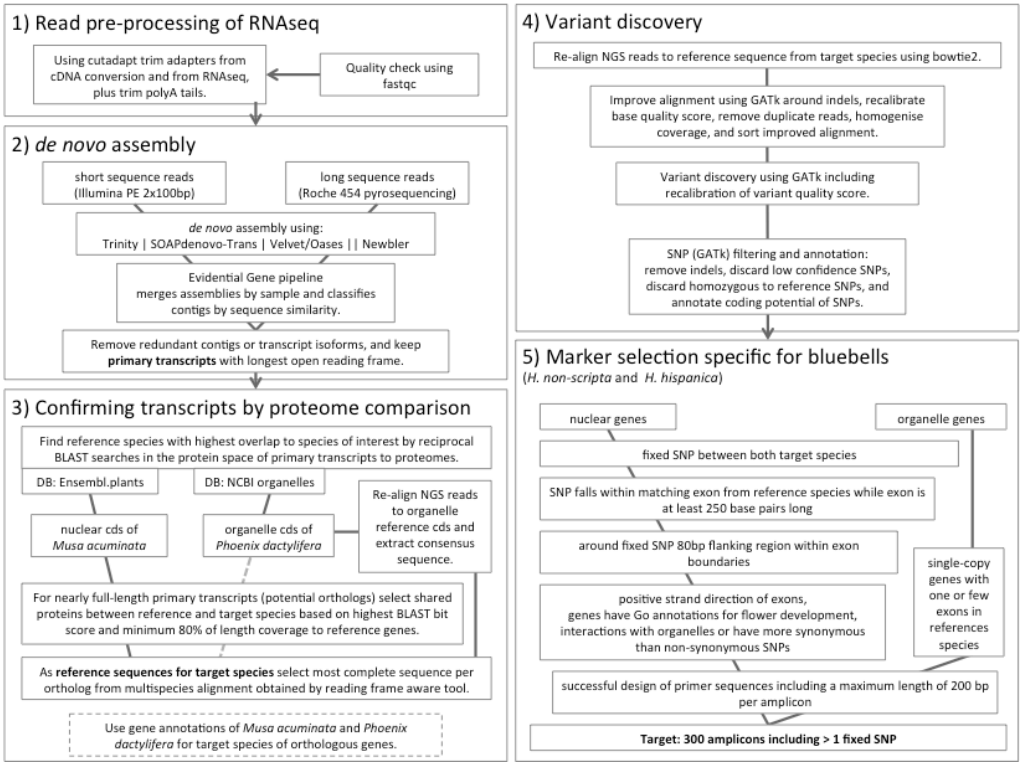


Figure 3.2 – Workflow of bioinformatics pipeline.

3.2.3 mRNA data and read pre-processing

RNA-seq libraries of three Illumina 2x100bp paired-end HiSeq2000 runs included the accessions BB-339 (*H. hispanica*, referred to as library SWA1), BB-411 (*H. non-scripta*, referred to as library SWA2), and BB-130 (*H. paivae*, referred to as library SWA3). The 454 Roche GS FLX presented a different genotype of *H. hispanica* (BB-356, referred to as library SWA4). Further details on library preparation and sequencing were provided by Steve W. Ansell and Pete Hollingsworth (pers. comm.). Remnants of mRNA poly-A tails, and adapters from cDNA conversion and sequencing were trimmed using cutadapt (Martin, 2011) and fastx toolkit[1]. The read quality was examined for length distribution, GC content, duplicated sequences, and k-mer content using FastQC[2]. Broken pairs of the trimmed paired-end libraries were filtered, but included as single reads in further downstream analyses, if suitable.

3.2.4 Individual transcriptome *de novo* assemblies

The highest possible number of different transcripts and their isoforms was obtained by using four assemblers with different k-mer sizes and merging the resulting contigs

[1] http://hannonlab.cshl.edu/fastx_toolkit/
[2] http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

based on sequence similarity and open reading frame. Each Illumina library (SWA1-3) was individually assembled using the pipelines of SOAPdenovo-TRANS v1.03 (Xie et al., 2014), Trinity release_2013-08-14 (Haas et al., 2013), and Velvet v1.2.10/Oases v0.2.8 (Schulz et al., 2012; Zerbino and Birney, 2008). The SWA4 library was assembled using Trinity, SOAPdenovo-TRANS and Newbler v2.6.0 (Roche, Mississauga, ON) on iplant (Goff et al., 2011). Before Trinity assembly with k-mer size 25, the read data was reduced using its k-mer in-silico normalisation. VelvetOptimiser v2.2.5 was applied[3] prior to the velvet pipeline, which optimises the primary parameters of k-mer size and expected coverage. The best support was given for k-mer 31 and thus only one run of Velvet was executed. SOAPdenovo-TRANS was used to assemble k-mer sizes of 25, 31, 75 for libraries SWA1-3, and 31, 63, 127 for the longer 454 reads of library SWA4. In addition, 454 reads were used as guidance only in SOAPdenovo-TRANS assemblies of the Illumina reads.

3.2.5 Joining assemblies and selecting set of 'best' mRNA sequences

Each species' raw assemblies were merged and run through the EvidentialGene tr2aacds pipeline (Gilbert, 2013) to remove redundancy and to select the 'best' mRNA sequence based on longest open reading frame (ORF). The pipeline translates the raw assemblies into all possible coding sequences (cds) based on continuous ORF and their amino acid sequences (aa). It removes redundant sequences employing exonorate/fastanrdb while protein qualities decide among identical cds which to choose (Slater and Birney, 2005), and contigs which are perfect fragments of longer cds are removed by employing CD-HIT-EST (Li and Godzik, 2006). The remaining non-redundant contigs are blasted against each other to find high-identity exon-sized local alignments within the cds. Based on this similarity score and protein quality, they are sorted into classes of primary or alternative transcripts (potential isoforms of primary transcripts), with okay and drop subsets. The primary-okay transcripts, which are the best mRNA sequences with longest open reading frame, were used in species comparisons to find shared orthologs.

The NGS reads were re-aligned to their primary transcripts using bowtie2 v2.1.0 (Langmead and Salzberg, 2012) and Trinity's alignReads.pl wrapper script (Haas et al., 2013). Alignment rates were taken from its output. To obtain counts of effective coverage, the alignment was adjusted around indels, the base quality re-calibrated and the maximum coverage set to 250 reads using tools from the GenomeAnalysisToolkit (McKenna et al., 2010), and finally duplicate reads removed using samtools' rmdup (Li, 2011). The average coverage per locus was counted in R version 3.2.1 (2015-06-18; R Core Team, 2016).

3.2.6 Proteome comparison and extracting homologs

To verify consistency of the *de novo* assembly, BLAST searches in protein space (BLASTp; e-value cut-off 1e-05; one maximum target sequence) were performed between amino acid sequences of primary transcripts of each bluebell and nine proteomes obtained from Ensemble Plants (release 77 Oct 2014; Flicek et al., 2014). The nine accessions/taxa were *Brachypodium distachyon* (2010-02-Brachy1.2), *Hordeum vulgare* (082214v1), *Musa acuminata* (MA1 (2012-08-Cirad)), *Oryza brachyantha* (OGEv1.4), *Oryza glumaepatula*

[3]https://github.com/Victorian-Bioinformatics-Consortium/VelvetOptimiser.git

(ALNU02000000 (2013-09)), *Oryza sativa* Japonica (IRGSP-1.0), *Setaria italica* (JGIv2.1), *Sorghum bicolor* (2007-12-JGI), and *Triticum aestivum* (IWGSC1.0+popseq (2.2)). The BLAST analysis was further annotated using the 'analyze_BLASTPlus_topHit_coverage.pl' script from Trinity (Haas et al., 2013). This script examines the percentage of the query sequence being aligned to by the best transcript, whereby the hits are reduced to unique results of the highest BLAST bit score and longest match length per transcript. All BLAST searches were performed using BLAST+ v.2.2.29 (Camacho et al., 2009). The species with the highest number of best matches to the bluebell transcripts was used as reference for selecting homologous nuclear transcripts.

Two steps of local alignments were applied to identify shared nuclear genes between the three bluebells and the reference species. First, all bluebell transcripts were blasted against each other in protein space (BLASTp; e-value cut-off $1e-20$; one maximum target sequence). Second, each bluebell species' transcripts were reciprocally blasted against the reference cDNA sequences (tBLASTn, BLASTx; e-value cut-off $1e-20$; one maximum target sequence) and analysed for completeness of the target gene. Finally, a custom python (v2.6.) script compared the unique best matches between the reference and the shared presence of the transcript in the three bluebell species. Those matches were restricted to a minimum of 80 % length coverage of the reference species' coding sequences as well as each bluebell transcripts to extract nearly full-length transcripts.

Chloroplast (cp) and mitochondrial (mt) genes from the bluebell RNAseq data, were obtained by either locally aligning the transcriptome coding sequences to the reference coding sequences, or directly aligning the trimmed RNAseq reads to the reference cds. As reference species the date palm, *Phoenix dactylifera* (cultivar Khalas, Al-Hasa Oasis, Saudi Arabia) was chosen because it is the closest relative that has both its organelle genomes completely sequenced (Fang et al., 2012; Yang et al., 2010). The coding sequences were downloaded from NCBI[4] (cp: NC_013991.2, mt: NC_016740.1).

The local alignment of the bluebell transcripts to the chloroplast and mitochondrion genes (tblastn; e-value cut-off $1e-20$; one maximum target sequence) recovered few genes, 33 % (of 96) and 54 % (of 43), respectively, but also multiple *de novo* transcripts matched these. By directly re-aligning the trimmed reads to the organelle coding regions of *P. dactylifera*, additional genes with at least 10 x coverage for two of three bluebells were obtained. Lower abundance of organelle mRNA in the RNAseq data (Gagliardi and Leaver, 1999) might be a reason for unassembled contigs of organelle genes. From the read alignments, consensus contigs were obtained for each bluebell using samtools, bcftools and its vcfutils.pl script.

For each marker system (cp, mt, and nuclear) the reference genes and the matching bluebell transcripts (obtained by *de novo* assembly or consensus sequence of read alignment) were extracted into individual alignments of orthologs. Coding frame aware sequence alignment was performed on all orthologs using prank v.140603 (Loytynoja and Goldman, 2005). To build a set of bluebell genes that will be used in the SNP calling as the reference, the longest complete transcript of either the Spanish or British bluebell was selected from each ortholog alignment. If the sequences were of the same length and without any insertions or deletions, the sequence of the British bluebell was automatically

[4]http://www.ncbi.nlm.nih.gov/nuccore/

preferred. If sequences of either the Spanish or British bluebell obtained larger sequence gaps (due to insertions, deletions, missing or ambiguous data), the alignment was checked manually and the most complete sequence chosen. From here onwards, this set of orthologous sequences will be referred to as the bluebell reference. The bluebell reference was double-checked for the longest open reading frame using transdecoder[5] (Haas et al., 2013).

3.2.7 Discovery of inter-specific genetic polymorphisms

Illumina reads were realigned to the bluebell reference using bowtie2/2.1.0 (Langmead and Salzberg, 2012) and Trinity's alignReads.pl wrapper script (Haas et al., 2013) for each species separately. Variant calling was performed following GATK Best Practices recommendations for RNA-seq (DePristo et al., 2011; Van der Auwera et al., 2013). The *HaplotypeCaller* was used for variant discovery between all three species to gain more confidence in variants and to potentially use the third species, *H. paivae*, as outgroup species to decide on the fixed target SNPs. GATK's variant recalibration includes a machine learning method to assign a probability to each variant in a raw call set. This variant quality score can be used in a second step to filter the raw call set, thus producing a subset of calls with our desired level of quality, fine-tuned to balance specificity and sensitivity. In order to use the *VariantRecalibrator* and *ApplyRecalibration* tools the initial variant call set was restricted to high quality variants (phred > 500, read depth > 10) and used them as positive training set of known variants in the recalibration. Final variant filtration was then based on the posterior variant quality score at the filtering tranche of a type I error of 99 %. Hence, sensitivity over specificity was chosen to discover all possible true variants at the cost of introducing more false positive variants. Lastly, all PASS single nucleotide polymorphisms were selected (*VariantFiltration*, *SelectVariants*; McKenna et al., 2010).

Variant discovery for the organellar genes was more challenging. In library preparation, poly(A) tails of the mRNA are captured for reverse transcription into cDNA. However, poly-adenylation can initiate mRNA degradation in plants (Gagliardi and Leaver, 1999) and makes organelle mRNAs prone to incompleteness. In addition, they often contain RNA editing sites (Fang et al., 2012; Meng et al., 2010). In the organelle genomes mostly RNA editing sites occur at C-U positions, which might be occurring as C-T sites in RNAseq data (Meng et al., 2010). Consequently, both haploid and diploid variant discovery were applied on the organelle reference genes to account for potential RNA editing sites and sequencing errors. In the end, as a conservative approach the diploid variant discovery was used and all heterozygous sites were removed. Hard filtering was applied because variant recalibration was not applicable for such low number of variants. SNPs that were discovered from less than five reads, or showed a low quality score (Phred Q < 50) were filtered.

The discovered SNPs were annotated for their coding potential using a customised python script. The variant position was compared with the position of the concordant amino acid sequence of the bluebell reference and whether the alternative allele would

[5]https://transdecoder.github.io/

alter the amino acid. The codon key was based on the generally expected nucleotide triplet coding for an amino acid for plants.

3.2.8 Selecting species diagnostic target regions

After variant discovery using three species, the data were restricted to SNPs with a quality score above 33, a minimum depth of five reads, and which occurred only between *H. non-scripta* and *H. hispanica*. To design markers that distinguish both species, a target SNP was designed to be homozygous for different alleles (i.e. AA vs BB). Since re-sequencing was planned from whole genomic DNA, where exons are interspersed with intron sequences, exon-intron boundaries as well as exons shorter than 160 bp were avoided and identified by local alignment (tblastx) of exon sequences from the reference species (obtained from Ensembl plants) to the bluebell reference. Potential target SNPs were selected that occurred within matching exon boundaries and providing 80 base pairs (bp) flanking region to either side of the SNP. This flanking region is required to fit primer oligos of 20 – 22 bp length within exon length. Loci where the reference species' exons aligned in reverse sequence order (negative strand) were removed. Those genes, which have gene ontologies (GO) annotations in the reference species related to flower development and cytonuclear interactions were prioritised (supplement 3.7.2). Primers for the target SNP and its surrounding nucleotides within potential exons were predicted around any other present SNP using primer3 (Untergasser et al., 2012). Primers were constrained to a maximum length of the amplicon (i.e. PCR product that includes the primer sequences) of 200 bp. The parameters (Tm: min: 59.0, opt: 60.0, max: 61.0; max poly-X: 3) were chosen to optimise a uniform amplification of the target sequences. For the chloroplast and mitochondrial genes the target SNP selection within exons was done manually because of the very small number of variant sites between both species. Preference was given to single-copy genes, and genes with only one or a few exons based on *Phoenix dactylifera* gene structure annotations (Fang et al., 2012; Khan et al., 2011).

3.2.9 DNA extraction for re-sequencing, Fluidigm and MiSeq

Fluidigm's Access Array system requires very pure extracts of high molecular weight DNA for the PCR reactions. Isolating DNA of bulbous monocots, such as species from *Hyacinthoides* was difficult due to the high concentration of carbohydrates and other secondary compounds especially in bulbs but also the leaves (Brocklebank and Hendry, 1989). At the same time, a high-throughput process for a population genetics study was intended. Therefore, a few steps from mainly three protocols/technologies were combined: Wang (2013) (TNE wash), Doyle and Doyle (1987) (CTAB lysis and SEVAC purification (Schneider et al., 2004; Trewick et al., 2002)), and Qiagen's BioSprint 96 DNA Plant Kit (Qiagen) and Qiagen's supplementary protocol for purification of PCR products using the BioSprint 96 workstation. For a detailed protocol see supplement 3.7.1.

The primer pooling in sets of 10s was optimised to avoid multiple target regions from the same gene per reaction well of the access array; and to avoid mixing organelle targets with nuclear regions due to expected varying genomic DNA template amounts. The access array was loaded ideally with 75 ng per sample. After PCR, the amplicons were pooled

across columns and harvested by individual to barcode every sample with a unique identifier. The amplicons were quantified, multiplexed and prepared for Illumina's mid length sequencing (MiSeq, 250 bp, paired end) on one lane. The MiSeq output was de-multiplexed by means of the individual barcodes. The primer design and pooling, PCR amplification and library preparation using the 48.48 Access Array were done by the Genome Centre at Barts and The London School of Medicine and Dentistry, UK.

3.2.10 Evaluating re-sequencing success

Illumina adapters and amplicon primers were trimmed using Trimmomatic v0.33 (Bolger et al., 2014) and cutadapt v1.8.1 (Martin, 2011). Especially, the palindrome method of Trimmomatic was applied to filter paired end reads that do not have both primer pairs. The read length is generally longer than the amplicon size and therefore good PCR templates should have both primer ends present. The primers and trailing bases with an average quality below 30 (default: across 4 bases) were cropped and subsequently read pairs shorter than 100 bp were removed because the minimum expected length of the target sequence was 105 bp. The (raw) reads' quality was visually assessed using FastQC (Babraham Institute, Cambridge, UK, 2011). All samples' reads were aligned to the target subset of bluebell reference genes using Trinity's (Haas et al., 2013) wrapper script alignReads.pl and bowtie2 (Langmead and Salzberg, 2012). Variant discovery and genotype calling was also done using GATK tools (McKenna et al., 2010) as previously stated, but using the best practise guide for DNA sequence data (DePristo et al., 2011; Van der Auwera et al., 2013). Read alignment was optimised around potential indels applying GATK tools *RealignerTargetCreator*, and *IndelRealigner*. The read's base quality was recalibrated using the *UnifiedGenotyper*, *BaseRecalibrator*, and *PrintReads* tools. For variant discovery the *HaplotypeCaller* was applied in the diploid (nuclear genes) or haploid (organelle genes) ploidy setting. Reads with a mapping quality below 44 and reads that matched several reference regions were masked, and also bases with a low quality score (Phred Q < 20) were ignored. When *HaplotypeCaller* is run in the variant discovery mode, it searches for variant blocks within each sample by locally realigning the reads into most likely haplotypes. By this, it provides likelihood estimates of observing alternative alleles at each position of the reference. Subsequently, using the *GenotypeGVCFs* tools, these haplotypes can be easily and time efficiently joined into multi-sample vcf files. Variant recalibration was used with a high quality subset of the just called variants as positive training sites. In addition, the target variants and the surrounding transcriptome variants were given as input for the confidence model of the recalibration. For the sake of specificity over sensitivity, a more stringent tranche of 90 % for the type I error of false positive discovery was selected. Lastly, hard filtering was applied to exclude variants, which are homozygous for the alternative allele in all samples (i.e. non-variant but to the reference sequence), and to exclude variants with missing information in more than 30 % of the samples.

In case of the organelle genes, hard-filtering excluded variants that presented fewer than 10 reads coverage, a Phred Q score below 30, and strand bias above 30.

Several tools were used to extract information or to reformat data formats, such as bamtools, samtools, bedtools, and vcftools (Barnett et al., 2011; Danecek et al., 2011; Li

et al., 2009; Quinlan and Hall, 2010). For more details on the applied parameters and mentioned python or R scripts see GitHub repository of the bluebell project[6].

3.2.11 Genetic clustering analyses of nuclear biSNPs

The re-sequenced data (biSNPs) was tested for the level of clustering that can be achieved between individuals and how effective the species identification by diagnostic alleles is. In addition, the genotype of plants growing in the CPG over years were compared to *in situ* samples. Therefore, genotype information was extracted for all bi-allelic single nucleotide polymorphisms and passed samples. The allele counts were transformed so that in each position '0' denotes the major allele across all samples, and '1' and '2' are the counts of the minor allele in each sample. To estimate allele frequencies (AF) between both species a folded joint site frequency spectrum was used – folded because the data was not polarised (i.e. knowing which allele is ancestral or specific for which species). The frequencies were categorised as follows: private alleles are AF = 0 in one species and any AF > 0 in the other, which includes fixed polymorphisms that occur at AF $\geq$ 0.95 for the other species, and lastly shared polymorphisms range between AF > .05 and AF < 0.95 for either species. For the organelle data genotype data were constructed where '0' denotes the non-scripta haplotype and '1' the alternative haplotype. For data manipulation and analyses R version 3.2.1 was used (2015-06-18; R Core Team, 2016).

A principal component analysis (PCA) of individual genotypes was conducted using *prcomp()* (stats package; R Core Team, 2016) with missing data imputed prior to analysis (R package MissMDA; Josse and Husson, 2016).

Bayesian clustering was performed with fastSTRUCTURE (Raj et al., 2014), and the number of potential ancestral clusters from K = 1 – 3 was tested. Data input was converted using PGDspider2 from vcf format (Lischer and Excoffier, 2012). The output was compared by log likelihood values for the most probable model applying the provided chooseK.py script.

Hierarchical clustering of individuals was performed based on the minimum hierarchical variance (Ward's 1963 criterion; R stats package; Murtagh and Legendre, 2014; Ward Jr., 1963) calculated from pairwise absolute distance between each locus (so-called Manhattan distance, R stats package). Clustering was depicted as a dendrogram (R package ggdendro; de Vries and Ripley, 2016).

3.3 Results

3.3.1 *De novo* transcriptome assembly statistics

After adapter trimming, about 30 million read pairs were retained for each Illumina library (SWA1, SWA2, SWA3) and 1 million reads from the '454' library (SWA4, Table 3.1). The individual assemblies resulted in 20,000 to 190,000 different contigs (Figure 3.3, Table 3.2). Trinity and Velvet generated the longest assembled raw contigs, evidenced by highest N50 (1,200 – 1,500 bp) and average lengths (668 – 1,000 bp). In contrast, SOAPdenovo-Trans generated some very long contigs (up to 40,111 nucleotides) but

[6]https://github.com/JeannineM/MarkerDev/blob/master/SupportingScriptNotesGitHub.Rmd

overall shorter average length contigs and lower N50 (Table 3.2). All possible transcripts, without much focus on the performance of a specific tool, were merged by library. Comparing their sequence similarity and locally aligning the sequences to each other showed that about 42.5 – 49.7 % of the contigs were redundant. Redundant contigs were fragments of longer contigs with 100 % sequence similarity (Figure 3.4). Non-redundant contigs were classified as primary transcripts (i.e. transcripts with the longest open reading frame) in contrast to their potential alternative transcript (i.e. transcripts of partial length of its primary transcript or alternative order of exons). SWA4 transcripts, from 454 reads, showed the highest fraction of primary transcripts compared to the number of obtained non-redundant contigs (12.7 %), while the other libraries ranged from 5.0 – 5.3 %. However, in absolute counts, twice as many primary transcripts were obtained from the Illumina libraries (about 32,000) than from the '454' library (16,000). The primary transcripts from the Illumina libraries showed good coverage ($\geq$ 10 x) for about 20,000 primary transcripts with about 64 % of the reads aligned, in contrast to the SWA4 transcripts that provided much lower counts (Table 3.3).

By aligning the RNAseq reads directly to the chloroplast and mitochondrion genes of *Phoenix dactylifera* and extracting the consensus sequence, 86 (of 96, 89.6 %) and 40 (of 43, 93 %) genes were obtained, respectively.

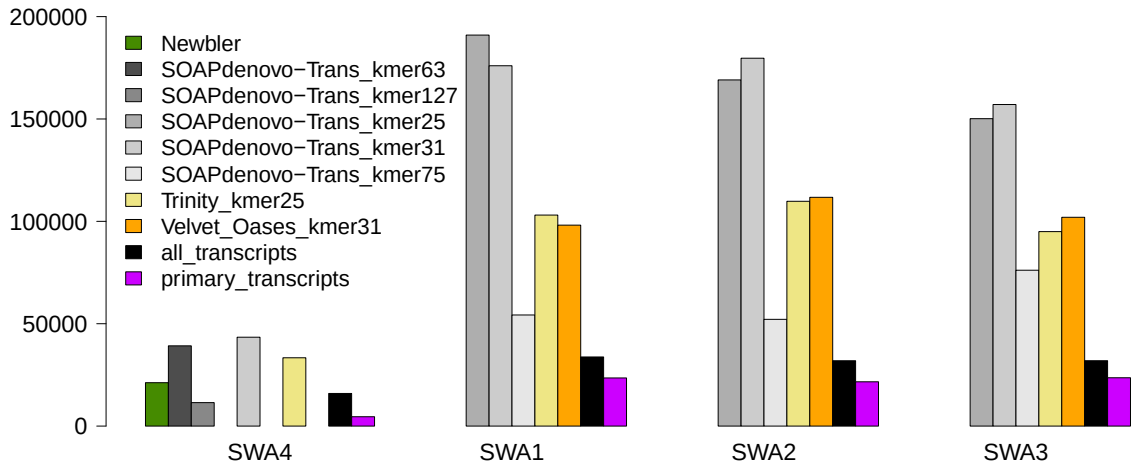


Figure 3.3 – Counts of raw contigs by different assemblers for each NGS library (SWA1-4). After merging contigs from different assemblers and removing redundant contigs, the total number of transcripts (black) and their subset of primary transcripts (purple) are also shown for each library.

3.3.2 Transcript identification and shared orthologous genes between the three bluebells and the reference species

Of about 32,000 and 16,000 primary transcripts of the Illumina and '454' libraries, about 66 % and 79 % respectively returned at least one BLASTp hit to one of the nine species. Based on highest bit score and longest aligned region between query sequence to data base, *Musa acuminata* showed the highest number of top BLAST hits (41 % for SWA1-3, and 52 % for SWA4, Figure 3.5). The *Musa acuminata* genome (D'Hont et al., 2012) was therefore used as reference species for a new BLASTp proteome comparison.

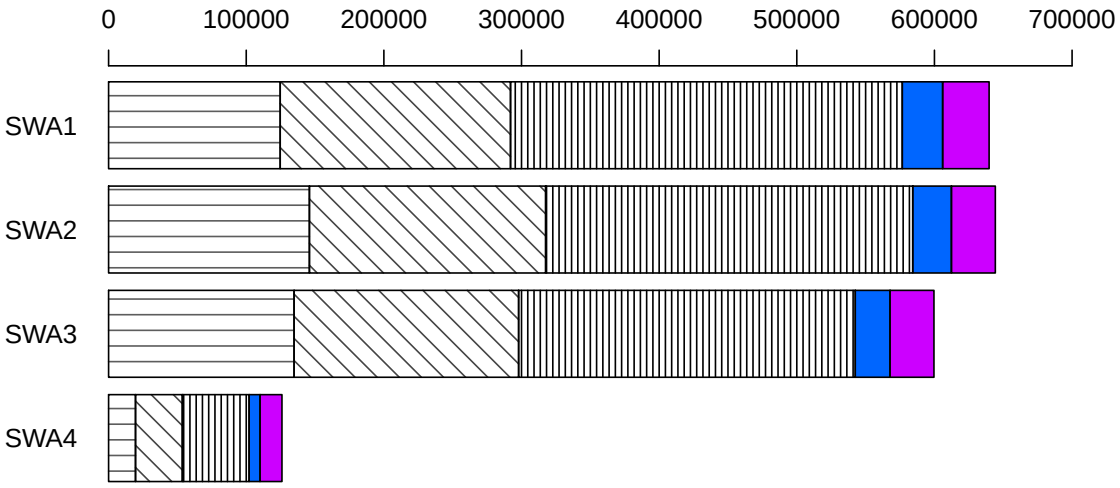


Figure 3.4 – Classification of longest open reading frame (ORF) coding sequences based on sequence similarity and length for each library (SWA1-4); annotated with: black patterned – redundant contigs (from left to right: 100 % identical cds, fragments of other cds, lower identity cds i.e. potential paralogues), blue – alternative cds (e.g. isoforms, splice variants, incomplete fragments), purple – primary transcripts (i.e. good support cds with longest ORF).

From a total of 36,519 banana cDNA sequences 23,118 (63.3 %) are met by bluebell transcripts in reciprocal blast searches. But only 5,968 are shared between all four libraries. Restricting the matches to a minimum gene coverage of 80 % length of banana genes only a small number of transcript matches remained (less than 500 transcripts). Since the SWA4 transcripts had the lowest count of unique matches to *Musa acuminata*, it limited our possible outcome and was therefore excluded from the comparison. Between the three Illumina transcriptomes, 1,047 genes that presented an alignment over at least 80 % length of a banana gene were shared (Figure 3.6(a)). The sequence similarity between bluebell query and banana protein ranged from 28 – 96 % (pident) with a mean of 71 %. The sequences of these contigs were extracted and aligned for inspection of the most complete sequence. The manual inspection of the ortholog alignments (including the sequences of *Musa acuminata* and the three bluebell species) showed one inconsistent gene that was then removed. From these alignments of 1,046 genes, the longest transcript sequence (with ORF) of either *H. non-scripta* or *H. hispanica* were extracted, which constituted our 'bluebell reference' of nuclear transcripts. Gene ontology terms and descriptions for the associated genes were used from the annotations of *Musa acuminata*'s proteome.

For the organelle marker, read alignment recovered bluebell sequences of 89 plastid and 40 mitochondrial protein-coding sequences in *P. dactylifera*. For both, this was a high proportion of possible genes: 92.7 % and 93 %, respectively. After extracting the consensus sequences for each sample and analysing the alignments for completeness, 86 plastid and 40 mitochondrion genes were kept as 'bluebell reference' of organellar transcripts.

For more details on the genes, which were found jointly between all three bluebell species and are further used as the bluebell reference (1,046 nuclear, 86 chloroplastic, and 40 mitochondrial genes) see Tables 3.5 – 3.7.

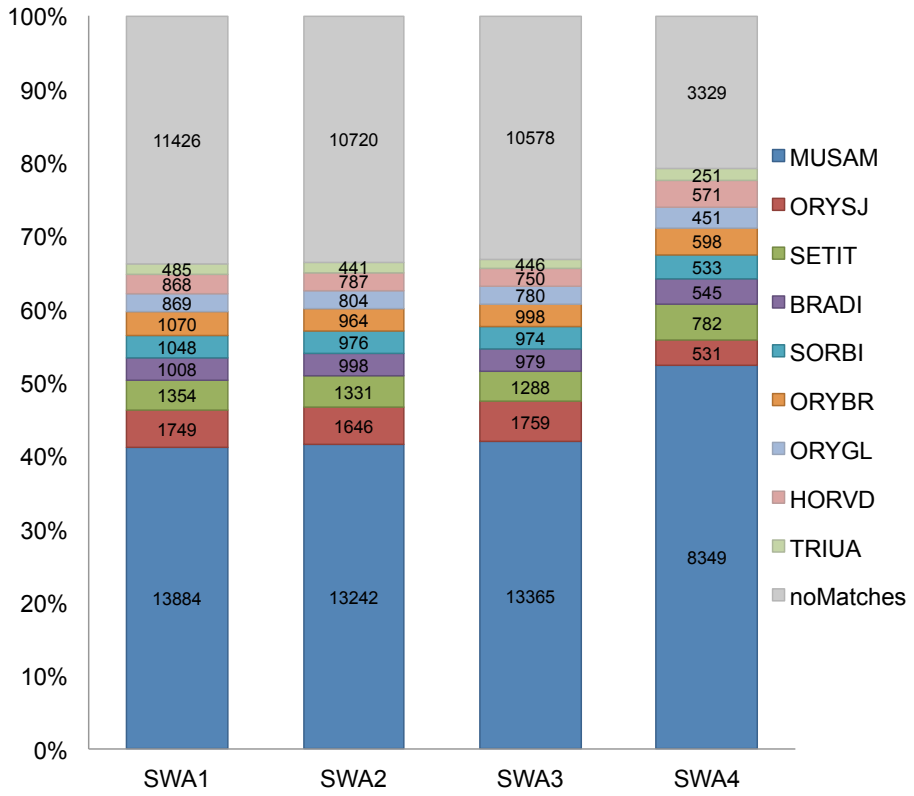
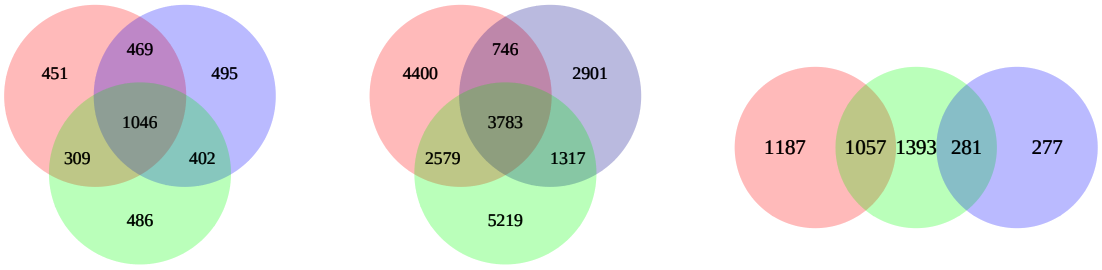


Figure 3.5 – Counts of best blast hits of bluebells' primary transcripts (libraries SWA1-4) to proteomes of nine species. Proteome species are abbreviated as: MUSAM – *Musa acuminata*, ORYSJ – *Oryza sativa* Japonica, SETIT – *Setaria italica*, BRADI – *Brachypodium distachyon*, SORBI – *Sorghum bicolor*, ORYBR – *Oryza brachyantha*, ORYGL – *Oryza glumaepatula*, HORVD – *Hordeum vulgare*, TRIUA – *Triticum aestivum*.

3.3.3 Genetic variability in 1046 shared genes for three bluebells

Amongst the nuclear bluebell reference transcripts, one was monomorphic but the other 1,045 presented a total of 20,945 single nucleotide polymorphisms (SNP) over 1.5 million nucleotides when calling variants for the three bluebell species. The transition:transversion ratio was 1.86. The third species, *H. paivae* was included to obtain larger confidence in the variant discovery. Out of all SNPs, 3783 (18.1 %) variant positions were shared by all three species (Figure 3.6(b)). *H. paivae* had the most private variants (5,219) and shared more of its other variants with *H. hispanica* (20 %) than with *H. non-scripta* (10.2 %), while *H. non-scripta* shared more of its polymorphisms with *H. paivae* (15.1 %) than with *H. hispanica* (8.5 %). *H. non-scripta* had the lowest number of total SNPs (*H. non-scripta*: 8,747, *H. hispanica*: 11,508, *H. paivae*: 12,898) but the highest number of heterozygous sites (*H. non-scripta*: 7,599, *H. hispanica*: 6,840, *H. paivae*: 7,210). Among the strictly homozygous variants (which are present in a subset of 805 genes), *H. hispanica* and *H. non-scripta* share no alleles, while *H. paivae* seems to bridge both species with the largest proportion of alleles shared with *H. hispanica* (Figure 3.6(c)). The low number of homozygous alternative alleles (and therefore overall lower number of polymorphisms) for *H. non-scripta* resulted from the fact that the bluebell reference sequences were mostly taken from library SWA2, i.e. *H. non-scripta*.

Regarding the organelle SNP calling, the diploid variant calling (cp: 207 SNPs, mt:



(a) Shared genes (b) Shared SNPs (c) Shared homozygous SNPs

Figure 3.6 – Venn diagrams of: (a) Shared genes between the three bluebells with over 80 % length coverage to *Musa acuminata* reference genes. (b) Number of variant sites polymorphic only in either species or polymorphisms shared between bluebells in a subset of 1.045 polymorphic genes. (c) Number of homozygous bi-allelic SNPs that are unique to each species or shared between species in a subset of 805 remaining genes. The species are colour coded with red – *H. hispanica*, blue – *H. non-scripta*, green – *H. paivae*.

385 SNPs) returned more variants than the haploid calling (cp: 130 SNPs, mt: 115 SNPs). For the diploid variant calling, the majority of heterozygous variants mostly occurred as C-T/T-C transitions (cp: 49.4 %, mt: 93 %), which are common RNA editing sites in plant mRNAs and were therefore removed. The final set of homozygous SNPs between *H. non-scripta* and *H. hispanica* for the plastid genome accumulated 120 SNPs in 41 genes totalling 79,616 nucleotides. The remaining 45 genes were monomorphic. For the mitochondrion genes there were 10 loci, which exhibit 13 SNPs over a total length of 33,372 nucleotides. The other 27 genes were monomorphic.

3.3.4 Applying filtering steps to select target regions

The marker design for re-sequencing was targeting homozygous SNPs for different alleles between *H. non-scripta* and *H. hispanica* ('fixed' sites). Between the two individuals 730 nuclear loci were found that contained 3,331 fixed sites. Using BLAST, the fixed sites were evaluated as to whether they occurred within an *M. acuminata* exon. Most of the variants fulfilled this criterion (727 genes with 3,315 SNPs). Next, a target SNP within an exon should have at least 80 bp genic flanking region on either side (581 SNPs in 275 different loci left). Lastly, the search was restricted to positive strand matches for exons to the bluebell reference and compared the remaining loci to selected genes with interesting GO annotations (153 of 1,046 genes). The intersection of these criteria resulted in 232 nuclear genes carrying 447 SNPs. Primer3 was applied to select oligos within a maximum amplicon length of 200 bp, yielding 222 different genes with 374 SNPs (in 290 target sequences).

Genetic variation in the organelles was lower and after restricting the screen to single-copy genes with sufficiently large exons to provide 80bp flanking on either side of a SNP, ten plastid and five mitochondrion genes were identified (sum 15 target sequences). Lastly five nuclear targets were dropped to generate 300 target sequences.

The final set of 300 target sequences (also referred to as amplicons) of the bluebell reference included 221 (361 SNPs) nuclear genes, ten (15 SNPs) chloroplast genes, and

five mitochondrial genes with one SNP each. Accordingly, this set of genes was referred to as the target genes. See Table 3.4 for an overview of the filtering steps.

The amplicons' length ranged from 150 to 200 bp (mean = 172.4 bp) including the primer sequences. In total, 39,480 nucleotides were expected in 300 fragments per sample from sequencing. These amplicons are expected to be fully covered by paired-end sequencing of 250 bp read length (using MiSeq).

3.3.5 Re-sequencing success of amplicons

The target amplicons were re-sequenced for 39 individuals of *H. non-scripta* (origin: Spain, France, UK) and 36 of *H. hispanica* (origin: Spain, Portugal) using multifluid multiplexing PCR and mid-length sequencing. The MiSeq reads were of good quality, observed by a peak at 34 for the average mean sequence quality. All the raw reads contained primer oligos, but no Illumina adapters from the barcoding or sequencing. Adapter trimming using the palindrome method (i.e. keeping reads containing both primers at their read ends) showed that on average 80 % of the reads contained both primers in their sequences and only 2.7 % of reads on average were removed because they failed other filters. As a result, about 35,000 reads per sample were aligned onto the 236 target genes. The mean alignment rate of trimmed reads exceeded 99 % per sample. A negligible fraction (95 of 5.2 M reads) of reads mapped onto multiple regions but was removed in downstream analyses. At least 21 nuclear amplicons (in 17 genes) were found that overlapped partially in their positions. Consequently, amplification failed for at least one of the overlapping amplicons - evidenced by few aligned reads (< 30 x). Another four genes (three nuclear, one chloroplastic) failed. The mean read coverage to remaining target genes per sample was high (128 x) and exceeded the minimum requirements (10 x) for variant calling. Noticeably, 91 % of the target sequences in 227 different reference genes were successfully amplified using multi-fluid-multiplexing PCR reactions.

3.3.6 Confirming the target SNPs

For ease of data analysis, structural variants were excluded while bi-allelic single nucleotide polymorphisms (biSNPs) were maintained. Bi-allelic SNPs were discovered for 214 nuclear genes from only 71 samples because two samples showed more than 12 % of missing data (samples 126-2, 126-9, probably due to fragmented gDNA input). Also, two other samples were removed, because they carried more than two alternative alleles at multiple target sequences (samples 262-B-01-CPG, and 353-02-CPG, potentially they are polyploids). In total, 1368 bi-allelic SNPs were discovered in 71 samples, of which 920 sites are new, and 448 are known SNPs from the transcriptome data. The latter include 263 of 361 (72.9 %) targeted SNPs, which were present in each of the 214 polymorphic nuclear genes. More interestingly though is their allele frequency (AF) and whether they are fixed or occur near fixation to provide diagnostic markers. For the majority of target SNPs (63.1 %) the alternative allele occurred at frequencies larger than 0 and below .95 of the samples for either species. Only nine fixed SNPs (AF = 0 and AF $>$ .95) were found. Looking for private alleles (AF = 0 in one, and AF $>$ 0 for the other taxon; i.e. including

fixed alleles) 94 SNPs were observed, for which the alternative allele is absent in species one, but present in species two at any rate.

Variant calling to organelles (haploid variant calling) included both samples from collecting site BB-126 because they provided enough read coverage, and the potential polyploid samples assuming maternal (haploid) inheritance of plastids. One plastid gene failed to amplify and another was monomorphic. The haploid variant calling discovered 19 variant sites in 13 different genes, of which one variant site was new. The variants can be grouped into three different haplotypes, whereby the chloroplast (14 sites in 8 genes) and mitochondrion markers (5 sites in 5 genes) were concordant (Figure 3.7). The British bluebell *H. non-scripta* exhibited one haplotype across its entire species range. The most common *H. hispanica* haplotype was different in 14 sites from the haplotype of *H. non-scripta* (Figure 3.7). A third organelle haplotype can be assigned to the two samples of *H. hispanica* from site BB-262, which had 5 SNPs in 4 different genes (3 mt, 1 cp) that were distinct from all other bluebell samples. In one other position of a chloroplast gene the two samples exhibited the reference variant, and otherwise they presented the *H. hispanica* haplotype (Figure 3.7).

3.3.7 Clustering of samples to test species assignment power of re-sequencing data

Since only a few fixed markers were recovered, it was tested if all biSNPs from re-sequencing allow species identification. Principal component (PC) analysis showed strong separation of both taxa along the first principal component, which explained 29.2 % of genetic variance in the observed data (Figure 3.8(a)). The second PC spread out intra-specific variation, especially the Spanish samples of *H. non-scripta* from the French and British samples; and both *H. hispanica* samples from Portugal, 188-B-31-CPG and 262-B-33-CPG (Figure 3.8(a) and 3.8(b)). For the third PC most samples were indistinguishable, except for the two *H. hispanica* Portuguese samples.

Similarly, Bayesian clustering presented the strongest support for two genetic clusters (Figure 3.9). Based on comparison of the marginal Likelihood estimates for K = 1-71, K = 2 presented the least negative estimate and therefore best fit to the data (marginal Likelihood = -0.5634 after 30 iterations). In addition, the chooseK.py script reported that two genetic clusters are sufficient to explain structure in the data. The posterior co-ancestry estimations per sample reported small proportions of admixture of *H. non-scripta* in *H. hispanica* (501-D – 6.8 %) and more frequently of *H. hispanica* in *H. non-scripta* (135-7 – 3.5 %, 391-3 – 4.1 %, 391-4 – 3.4 %, 398-5 – 1.45 %). These samples were collected *in situ*. For the Portuguese samples of *H. hispanica*, 188-B-31-CPG and 262-B-33-CPG, K = 3 revealed the samples contained around 9 % ancestry from a potential third cluster (data not shown, but appearing with admixed ancestry in Figure 3.9). The DNA of these Portuguese samples was derived from individuals maintained at the Chelsea Physics Garden and not *in situ*.

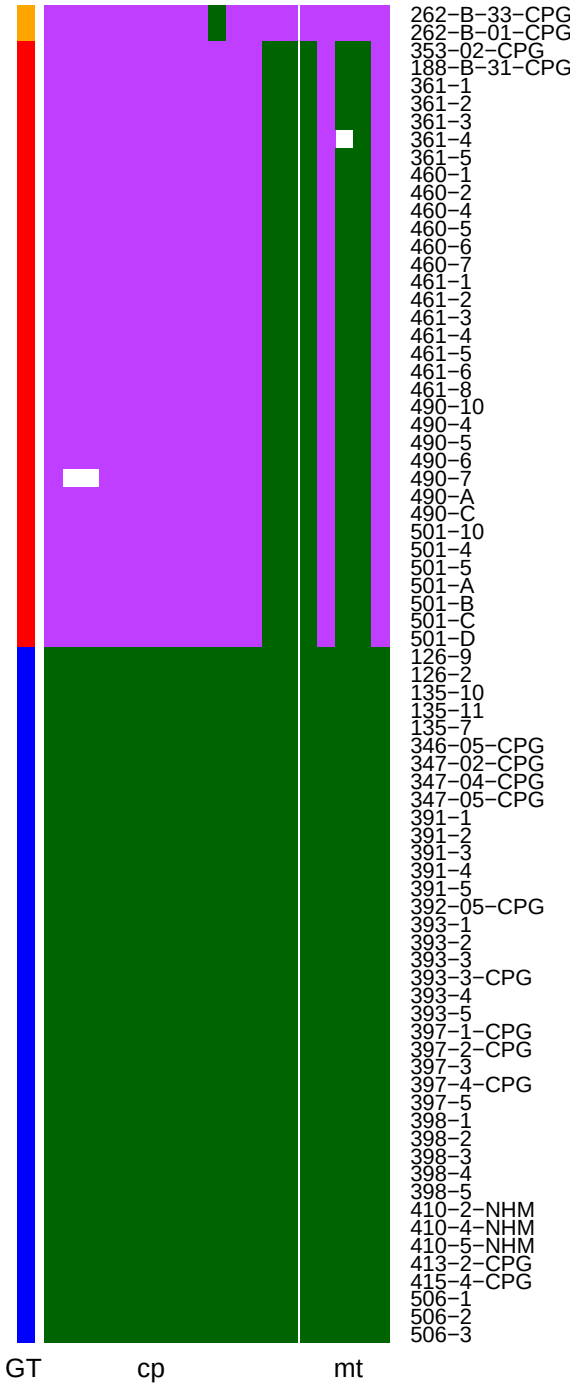
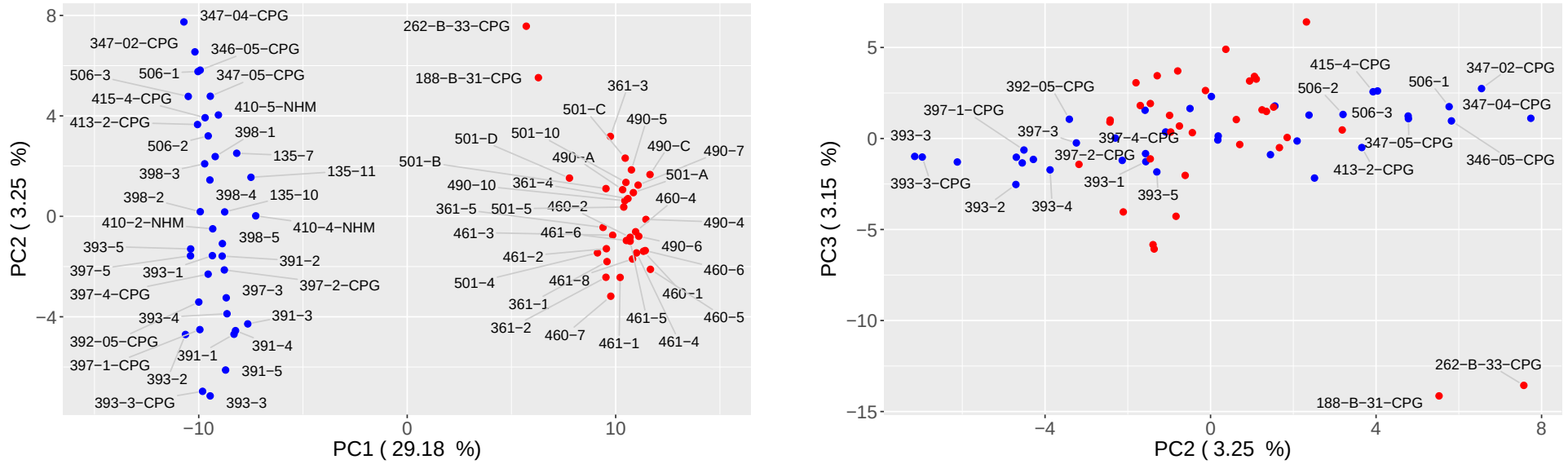


Figure 3.7 – SNPs in organelle genes (columns, green – reference allele, pink – alternative allele, white – missing data) are shown for 75 samples (rows). Left column (GT) shows the assigned haplotype with blue – *H. non-scripta*, red – the common *H. hispanica* haplotype, and orange the third haplotype for Portuguese samples from BB-262. The first data block represents 14 SNPs in eight chloroplast genes (cp) and to its right five SNPs in five mitochondrion genes (mt).



(a) Principal components 1 and 2

(b) Principal components 2 and 3

Figure 3.8 – Principal component analyses of nuclear biSNPs for 71 individual samples. Expected species identity is highlighted by blue – *H. non-scripta*, and red – *H. hispanica*.

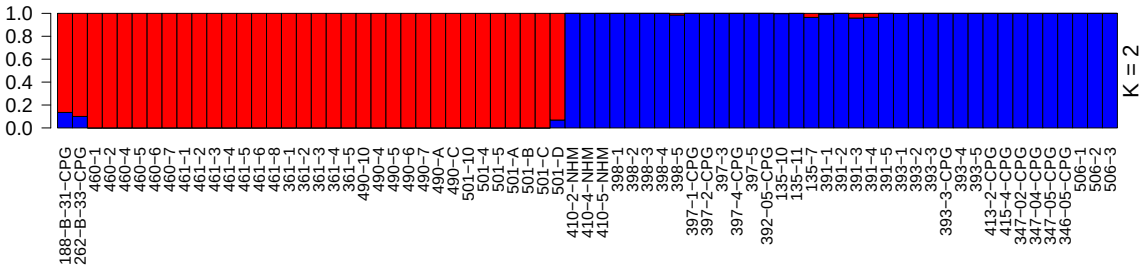


Figure 3.9 – Co-ancestry plots using 1368 nuclear biSNPs. Proportions of ancestry per sample for K = 2 with blue – *H. non-scripta* genome proportion, and red – *H. hispanica*.

Hierarchical clustering by pairwise Manhattan distance of the genotype frequency per sample mostly resolved genetic structure by collecting sites (Figure 3.10). Within *H. hispanica*, all individuals were grouped into their collecting sites. The two individuals from southern Portugal are placed as sister-group to all accessions of *H. hispanica*.

The clade of *H. non-scripta* individuals formed two major sub-clades (Figure 3.10): clade A represented individuals from a broad range including UK (BB-506), France (BB-346, 347, 413, 415) and northern Spain (sites BB-135, 410, 398), whilst clade B included individuals exclusively from the Cantabrian Mountains in Spain (BB-391, 392, 393, and 397), which occurred in close proximity. The individuals did not generally cluster into their collecting sites.

3.3.8 Identity of botanical garden plants using hierarchical clustering

Another interest of this study was to test if the 15 sampled individuals, which have been growing in the Chelsea Physics Garden, London, UK, since they were collected in Iberia in 2008, are still the original resource material. Potentially, new individuals that present patterns of introgression might have replaced the original individuals. Six *H. non-scripta* samples from France (in clade A) clustered with each other (346 with 3x 347; and 413 with 415). Unexpectedly, BB-413/415 (Brittany, Western France) were grouped with BB-398 (Galicia, Spain, non-CPG samples) instead of with BB-346/347 (Paris area, Northern France). The single replicate from the living material and the silica material, sample 393-3 (marked with '*' in Figure 3.10) could be confirmed genetically as almost identical (Manhattan distance = 0.01 with one SNP difference). The collecting site, BB-397, included three CPG samples and two silica samples. The BB-397 samples mostly cluster with each other, although they are interspersed with samples from BB-393, 391 and 392 (Figure 3.10). The two diploid Portuguese samples 188-B-31-CPG and 262-B-33-CPG clustered closest to each other. In addition, while BB-188 exhibited the common *H. hispanica* organelle haplotype, 262-B-33-CPG showed the third organelle haplotype. The second individual from this population, 262-B-01-CPG – although potentially polyploid – confirmed this haplotype. Genetic admixture, inferred from the fastSTRUCTURE, was not observed for the remaining CPG samples. Hence, introgression from *H. hispanica* into these *H. non-scripta* from other plants in the CPG is unlikely.

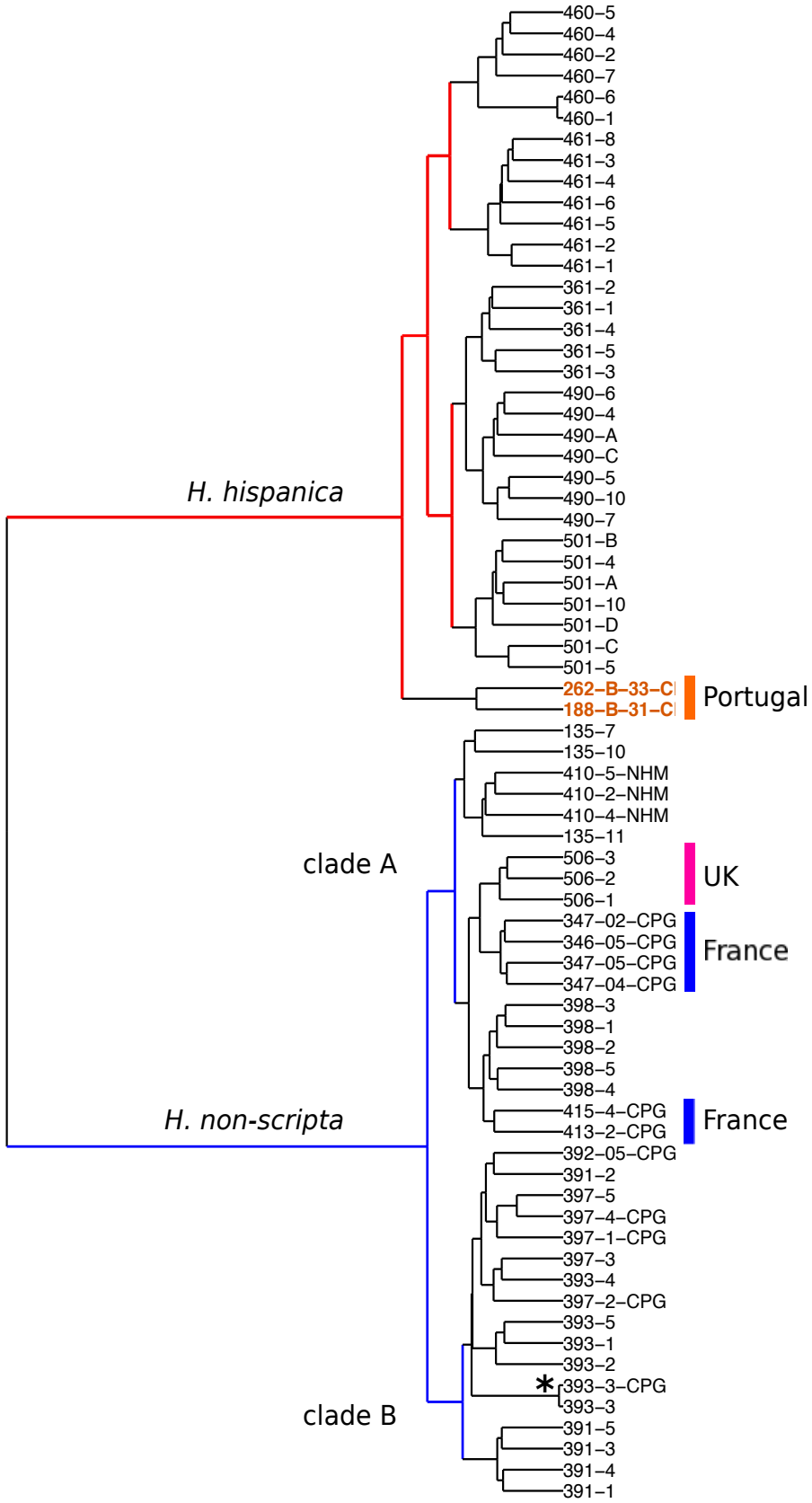


Figure 3.10 – Hierarchical clustering of 71 samples (four failed or potentially polyploid samples are missing) using pairwise Manhattan distances between samples' genotypes at nuclear genes. Highlighted are the samples with origins outside of Spain. The asterisk '*' denotes the replicated sample with two different sources of material.

3.4 Discussion

Next generation sequencing is promising access to genomic resources for populations of non-model organisms, and yet, for species with large genomes, many challenges remain to develop and apply population genetic markers (Egan et al., 2012; Sarah et al., 2016). Here, an approach using transcriptome sequencing of only four accessions was demonstrated to find genic regions with diagnostic properties for application in population genetic studies. The approach was tested to bluebell species of the genus *Hyacinthoides*, which have large genomes (1C > 22 GB) and strong conservation interest.

3.4.1 Illumina vs 454 sequencing

Many tools have been developed for *de novo* transcriptome assembly without a reference genome (Trinity, Mira, Velvet, SOAPdenovo-TRANS, CLCbio, Newbler), and best practice guides (De Wit et al., 2012; El-Metwally et al., 2013; Schliesky et al., 2012; Strickler et al., 2012) plus quality check tools (Li et al., 2014; Nakasugi et al., 2014) have led to an increasing number of published plant transcriptomes (Der et al., 2011; Fugate et al., 2014; Ness et al., 2011; Palma-Silva et al., 2016; Sarah et al., 2016). However, no single tool can reliably assemble all different/potential transcripts, not least because different assembly parameters, such as k-mer size, generate different outputs (Moreton et al., 2015). To obtain a large diversity of assembled transcripts, four different assemblers were employed and their contigs evaluated for sequence and structural similarity using the EvidentialGene pipeline (Gilbert, 2013). The EviGene pipeline was applied to obtain the 'best' transcripts based on longest open reading frames (Huylmans et al., 2016; Nakasugi et al., 2014; Visser et al., 2015). This removed much redundancy and resulted in about 30,000 potential primary transcripts from only one tissue, which is similar to other transcriptome characterisations of plants where multiple tissues were used (Ness et al., 2011; Palma-Silva et al., 2016; Sarah et al., 2016).

Each of the three Illumina HiSeq (PE 100bp) libraries provided more sequence information than the 454 Roche library, evidenced by number of *de novo* primary transcripts and their coverage by reads. In absolute counts the Illumina libraries provided more shared full-length transcripts between the three closely related bluebell species (1,046 genes). However, in relative proportions the assembled primary transcripts of 454 sequencing returned more matches to other proteomes (79 %) than HiSeq (66 %). In addition, the Illumina libraries produced a large proportion of alternative transcripts (like splice variants), and consequently about 40 % of contigs were not used in our pipeline. About 35 % of the reads did not align to the primary transcripts, and consequently failed to be assembled. For variant discovery it is common practise to restrict polymorphisms to regions with read coverage of $\geq$ 10 x. The effective mean coverage (after duplicate removal) for 454 sequencing was only 5.39 $\pm$ 9.75 x and only 1,554 transcripts had a coverage $\geq$ 10 x. This is in contrast to the short read Illumina libraries, which provided sufficient coverage (mean 33 – 38 x coverage) for at least two thirds of their primary transcripts (around 20,000) and enough reads to call high quality variants. Previously, 454 Roche platform was preferred for *de novo* assemblies because its longer reads better resolved difficult genomic regions like repetitive elements and alternative splice variants (Ekblom and Galindo,

2011; Ness et al., 2011). Even so, for marker development from transcriptome data distinguishing closely related species, the absolute number of shared genes and the discovery of high quality variants have to be prioritised. As a result, using Illumina sequencing for a limited amount of RNA accessions provides a 'good enough' transcriptome assembly for the purpose of marker development. Recent publications also preferably used strategies integrating short read sequencing with greater depth (e.g. Fugate et al., 2014; Guo et al., 2015; Li et al., 2014; Qi et al., 2016; Sarah et al., 2016; Steele et al., 2012).

3.4.2 Quality of the genes that were assembled and annotated

The comparison of *de novo* transcripts to proteomes allows the characterisation of gene function and enables a comparison with orthologous sequences available in public databases. Using reciprocal BLAST searches to nine proteomes from Ensembl plants, about 21,000 transcripts were identified (66 % per Illumina library) for each bluebell species. Previously, similar counts of shared reference transcripts to the SwissProt database were found from an exhaustive study of 26 transcriptomes (60 % with BLAST hits; Sarah et al., 2016). Comparatively few transcripts (1,046) were shared between all bluebell individuals and aligned with reference, orthologous proteins from *Musa* (Zingiberales) over more than 80 % length. The sequence similarity (pident) for these regions averaged 71 %, which is generally perceived as low. Several studies used higher sequence similarity to query orthology, but at lower length coverage (e.g. Sarah et al., 2016). Gene models and transcriptome annotation ideally require close reference species, but the most closely related nuclear proteome to bluebells (*Hyacinthoides*, Asparagales) available in late 2013 was *Musa acuminata* (Zingiberales), and for the organelles it was *Phoenix dactylifera* (Arecales). The divergence of the lineage including bluebells, bananas, and palms is estimated to date back into the Cretaceous (Hertweck et al., 2015). Thus, if more stringent sequence similarities parameters had been applied, it is likely that this would have resulted in fewer genes identified with highly conserved regions.

The approach did, however, risk the selection of markers resolving, for example paralogues, which was undesirable. Indeed, some regions were observed with multiple alleles in diploid samples in the re-sequencing data, indicating non-target amplification. Stringent mapping filters removed most of these reads. Other studies, especially when focussing on full transcriptome characterisation, rather than marker development, have also applied additional methods for identification of orthologous sequences, including contig clustering methods such as OrthoMCL (Fischer et al., 2011) and comparing the sequences to further databases like CEGMA, TrEMBL, or PFAM (e.g. Salgado et al., 2014; Vatanparast et al., 2016).

3.4.3 Exon/intron boundaries when re-sequencing from gDNA

Considerable thought was invested into selecting the target sequences around target SNP and any other variant present. Designing markers from exonic regions has advantages because of relatively low diversity and a reduced likelihood of primer binding to non-target sites (Harrison and Kidner, 2011). Nevertheless, re-sequencing these from genomic DNA can remain problematic. For instance, some studies claimed amplification failures due

to exon/intron boundaries (Sindhu et al., 2014). Primer sites that span such regions are undesirable, but are likely to occur given the short mean length of exons of 90 – 120 bp across eukaryotic genes (Deutsch and Long, 1999). The mean exon length of *Arabidopsis thaliana* and *Oryza sativa*, for which complete genomes are sequenced, are longer, 236.8 and 250.2 bp respectively (Koralewski and Krutovsky, 2011), although estimates vary (Kaplunovsky and Bolshoy, 2011; Zhu et al., 2009). To mitigate against that problem, exons from the reference species, *Musa acuminata*, were mapped onto the bluebell reference and amplicons shorter than 200 bp designed to avoid potential exon/intron boundaries. There was no failure of re-sequencing using paired-end MiSeq caused by introns.

3.4.4 Other pitfalls

The majority of amplicon failures occurred in the amplification step due to overlapping primer regions. This could have been evaded by separating primer pairs targeting the same exon region into different primer pools for Fluidigm's access array. Alternatively, the selection of primer pairs could have been restricted to one amplicon per exon. The primer specificity was high, although we noticed some potential paralogs (see discussion above). The samples were sequenced at greater depth than needed (mean 128 x) and there was potential to include more regions. Fluidigm can pool up to 480 primer pairs (i.e. ten primer pairs per reaction well[7]), but this project had only resources for 300 primer pairs.

Fluidigm is tested for high success rates of amplification in human DNA and demands an input of 7,500 genome copies (i.e. 1C = 3 pg; 50 ng DNA input diluted in up to 3 μl[7]). In large genome species, those quantities of high molecular weight DNA are challenging. Following instructions, the required input for bluebells with a genome size of 1C = 23 – 25 pg would have been 350 ng/sample, which was difficult to achieve. In addition, the DNA aliquots were increasingly 'gooey' at high concentrations, despite purification steps. Consequently, the array was typically loaded with only 75 ng/sample and the lowest DNA input was 60 ng. Other studies, using Fluidigm for PCR amplification, used similar amounts regardless of the sample's genome size (Uribe-Convers et al., 2016). However, high read counts, as a proxy for successful amplification, was not linked to DNA input. Instead, DNA fragmentation from poor or old silica material might have been a reason. Genomic DNA should be provided of high purity and high molecular weight, and then a lower number of gene copies per reaction compared with the recommended amount should be suitable for Fluidigm technologies.

Despite the potential pitfalls, amplification was successful in 91 % of the targeted sequences through high primer specificity, and the method showed good scalability to multiple samples. Similarly, Salgado et al. (2014) validated 90 % putative SNPs (172/191) using an allele-specific amplification strategy to rubber (*Hevea brasiliensis*, 1C = 2.11 pg), Fugate et al. (2014) validated 60 % of primer pairs (43/72) resulting in polymorphic SNPs that differentiate a set of eight sugar beet genotypes (*Beta vulgaris*, 1C = 1.25 pg), and (Ophir et al., 2014) achieved 71 % (346/480) informative SNPs from 105 accessions of pomegranate (*Punica granatum*, 1C = 0.72 pg) using Fluidigm – all species with small genomes by comparison to bluebells.

[7]http://www.fluidigm.com – manufacturer's instructions

3.4.5 Handling PCR amplicon data

The amplicon data were aligned to the bluebell transcriptome reference. Duplicate removal should not be applied because in most software read start- and end-positions are used to detect duplicate reads. But PCR based data should have the same start and end positions. Unfortunately, because duplicates were not removed, there are large alignment files and equivalent computational resources are required to store and effectively mine the data. In variant discovery, however, the primer regions, which can produce variants due to the primer sequence, should be removed, restricting analysis to the targeted sequences. In addition, absolute read depth should not be taken as a primary call for data confidence, because following PCR data, sequencing errors also amplify. Consequently, quality scores normalised by samples depth, read mapping quality, and tested for particular positions of a variant within a read were used as filters[8].

3.4.6 Evolutionary studies

Once the polymorphism data has been obtained, it can be used to estimate descriptive population genetics statistics and passed to tools, e.g. estimating coalescence history (Ellegren, 2014). Yet, in framing the research project one should be aware of the lack of additional information also due to missing reference species for non-model organisms, especially for monocotyledons. The obtained data, using our approach, is likely targeting conserved markers; there are no means of estimating recombination rates between markers (due to no linkage map of the physical distance between the target sequences); and the target sequences do not represent a random sampling of the species genome. The non-random design of markers can be replicated in simulation studies to account for potential biases for non-neutral marker properties.

3.4.7 Power to discriminate both bluebell species

As a proof of concept, diagnostic markers (more than 1000 SNPs) were developed to distinguish two closely related species to our satisfaction. The markers represent a reduced but expressed fraction of the large nuclear genome of both species, *H. non-scripta* and *H. hispanica*, and their organelle genomes. There is sufficient information to delimit both taxa, and to resolve intra-specific variation across the species' ranges.

Inference of co-ancestry showed potentially admixed samples between both taxa sampled from Northern Spain. They were collected at the margins of a known region of hybridisation along the central sierras of Spain (Grundmann et al., 2010). Therefore, applying this marker set to additional samples from this region seems promising to study hybridisation between both taxa.

For samples of *H. non-scripta* growing in a botanical garden (Chelsea Physics Garden London – CPG) for years, their origin was confirmed by comparison to other *in situ* diversity or replicated silica DNA material. However, we cannot exclude that there has been genetic mixture with the local garden diversity from *H. non-scripta* because SNP data failed to detect strong differences between *H. non-scripta* from UK, Spain, or France.

[8] https://software.broadinstitute.org/gatk/gatkdocs/

For *H. hispanica* from Spain the developed markers resolved local diversity. The four Portuguese *H. hispanica* samples are more complex to judge because their material comes from the Chelsea Physics Garden. Two of the samples appeared as potential polyploids (BB-262-B-01-CPG, BB-353-02-CPG) and need to be further analysed for their allele dosage and cytogenetics. Triploid individuals are known for *H. hispanica* and *H. non-scripta* (Grundmann et al., 2010), and also triploid hybrid individuals from the UK (Wilson, 1958). One exception in the genus is species *H. cedretorum* (Pomel) Rumsey occurring in the high mountain chains of northern Africa and is assumed a tetraploid offspring of *H. hispanica* (Grundmann et al., 2010). But no polyploid bluebells have been reported from the Iberian Peninsula (Grundmann et al. (2010); chapter 2). Potentially the additional alleles are a result of DNA contamination. The two diploid samples showed about 9 % admixture from unknown source.

The haploid organelle genome shows a third haplotype for samples of BB-262 and places BB-353-02-CPG and BB-188-B-31-CPG within *H. hispanica*. Without further sequencing of hybrid bluebells from UK and *H. hispanica* from Portugal any suggestion here would be speculation. However, Grundmann et al. (2010) analysed five non-coding chloroplast regions for a phylogenetic framework, in which the sample BB-262-01 (the original silica material of BB-262-B-01-CPG) formed a clade with sample BB-236-01 – '*H. hispanica* cultivar ("Spanish Bluebell") from Sydenham Hill, London', Figures 3 and 4. Therefore, the third organelle haplotype could be useful to determine maternal dispersal patterns of the so-called alien '*Spanish bluebell*' in the planned re-sequencing of British hybrid samples.

3.4.8 Failure to specify fixed markers

The sampling of genetic differences between two transcriptomes was not sufficient to discover fixed alleles and additionally could potentially lead to ascertainment bias, in which the allele frequencies from the two samples did not reflect the diversity in re-sampled population (Lachance and Tishkoff, 2013). But the large amount of shared polymorphism might also represent the close relationship between both species. Additional outgroup samples from the genus could have provided more information about the genetic diversity and state of ancestral alleles in both analyses – transcriptome and amplicon.

3.5 Conclusion

Based on three assembled bluebell transcriptomes, we established about 1,000 genes and primers for about 200 genes, which can easily be applied to wider data sets. This is envisioned to provide data for introgressive hybridisation between *H. non-scripta* and *H. hispanica* in the UK and northern Spain. Further, the bluebell transcriptomes present a resource for further studies, for instance a full transcriptome characterisation now remains practical.

The developed pipeline can be used as a guide for other population genetic studies of non-model organisms with large genomes. Scripts were provided on GitHub[9].

[9] /github.com/JeannineM/MarkerDev/blob/master/SupportingScriptNotesGitHub.Rmd

3.6 Contributions and acknowledgements

Thanks to the Genome Centre at Barts and the London School of Medicine and Dentistry, UK, for primer design and conducting the re-sequencing including library preparation using Fluidigm, and MiSeq sequencing. Additionally, I thank Prof Dirk Metzler (from University Munich) and Georgia Tsagkogeorga (from Queen Mary University of London) for fruitful discussions regarding the marker design and Alexandre Blanckaert (from University of Vienna) for help with coding python scripts.

3.7 Supplement information

3.7.1 Detailed DNA extraction protocol

Per sample 20 to 50 mg of dried leaf tissue were placed into 2 ml tubes with two glass beads and frozen over night in an -80°C freezer. The frozen material was grinded at 30 HZ for 2 x 1 minutes into a fine powder using a TissueLyzer II bead mill (Qiagen, Venlo, The Netherlands). Subsequently, the samples were stored on ice and 900 μl of ice-cold TNE buffer (200mM Tris-HCl, 250mM NaCl, 50mM EDTA) was added for pre-lysis washes. After dissolving the powder through vortexing, the cell suspension rested for 10 minutes on ice. The tubes were centrifuged for 5 minutes at 5000 rpm and the supernatant removed. This cleaning step of the cell suspension was necessary to remove mainly polysaccharides before cell lysis. This step was repeated several times until the supernatant became clear and liquid. For cell lysis 300 μl of 3 % CTAB buffer, 33.5 μl of 10 % Sarkosyl and 1.6 μl of Proteinase K (40 units/ml) were added to each tube, vortexed until cell suspension dissolved, and incubated in a heat block at 60 – 65°C for one hour. Approximately 2/3 of the volume of SEVAC (300 μl; 24:1 chloroform:isoamyl alcohol, (Schneider et al., 2004; Trewick et al., 2002)) were added to the lysate, mixed by inversion of the tubes and centrifuged for 3 minutes at maximum speed.

For the purification procedure, wells in S-Blocks and an elution plate were prepared as follows: slot 1 – 200 μl isopropanol and 20 μl MagAttract SuspensionG (Qiagen, Venlo, The Netherlands); slot 2 – 450 μl RPW buffer; slot 3 and 4 – 500 μl 100 % Ethanol each; slot 5 – 500 μl of 0.02 % (v/v) Tween 20 (Sigma-Aldrich, Gillingham, UK) in DNA grade water (Fisher Sciences, Loughborough, UK); slot 6 – elution in 200 μl of PCR grade water; and slot 7 – rod cover tip plate. The top aqueous layer (max. 400 μl) from the SEVAC purification was carefully transferred into the prepared S-Block (slot 1) and mixed with the isopropanol/MagAttract suspension by careful pipetting. After 10 minutes of incubation at room temperature the Biosprint 96 Plant program (Qiagen, Venlo, The Netherlands) was started.

The quality of the DNA extracts was examined by three methods: (1) The purity of all DNA extractions was measured using NanoDrop 8000 (TermoScientific Loughborough, UK); specifically by looking at the absorbance at 260 nm for DNA amount [ng/μl], at the A260/280 ratio (a ratio of 1.8 is considered 'pure' DNA[10]) and the A260/230 ratio, which can indicate contaminants in the extraction (Thermo Scientific, 2013). An A260/230 value in the range of 2.0 - 2.2 is referred to as normal, whereas much lower values may be the result of carbohydrate carry over, which is common for plants. Values smaller than 1.5 are likely to fail PCR due to residual chemical contamination from the extraction procedure. (2) Fragmentation of genomic DNA for all samples was explored by running 2 μl of the extract (loaded with 3 μl nucleic acid stain) on a 1.7 % agarose gel at 75 V for 60 minutes. The anticipated genomic DNA size was about 12 kb, anything below was considered fragmented or degraded. (3) The Qubit 2.0 Fluorometer (Life Technologies, Paisley, UK) was used with the dsDNA BR Assay to measure exact double stranded DNA concentrations. Per 96-well plate of extracts, 15 samples were measured using Qubit and

[10] nanodrop.com/Library/TO42-NanoDrop-Spectrophotometer-Nucleic-Acid-Purity-Ratios.pdf; accessed 20th November 2016.

the average ratio of NanoDrop (ND) and Qubit (Q) concentrations was applied to the remaining samples. A ND/Q ratio between 1 and 3 was considered a good measurement for pure extracts.

DNA extracts that failed these DNA quality requirements were additionally cleaned using the Qiagen clean-up protocol for the BioSprint96 robot. The S-blocks were prepared as follows: slot 1 – 200 μl extract, 200 μl PM buffer (Qiagen, Venlo, The Netherlands), and 20 μl MagAttract SuspensionG; slot 2 – 500 μl PE buffer (Qiagen, Venlo, The Netherlands); slot 3 – 500 μl 100 % Ethanol; slot 4 – 500 μl of 0.02 % (v/v) Tween20 in DNA grade water; slot 5 – elution in 100 μl of PCR grade water; and slot 6 – rod cover tip plate.

3.7.2 Selected gene ontology terms from *Musa acuminata* gene annotation

List of gene ontology terms selected for flowering related genes from 'biological process' keywords: anthocyanin-containing compound biosynthetic process, anthocyanin-containing compound metabolic process, aromatic amino acid family biosynthetic process, aromatic amino acid family metabolic process, carpel development, chiasma assembly, embryo development, embryo development ending in seed dormancy, embryo sac development, embryo sac egg cell differentiation, embryonic pattern specification, flavonoid biosynthetic process, floral organ formation, flower morphogenesis, growth, male meiosis, meiotic chromosome segregation, meiotic nuclear division, meristem development, meristem initiation, meristem maintenance, meristem structural organization, photoperiodism, flowering, petal formation, organ morphogenesis, ovule development, pollen development, pollen exine formation, pollen tube growth, pollen tube guidance, pollen wall assembly, polysaccharide biosynthetic process, polysaccharide catabolic process, primary shoot apical meristem specification, proanthocyanidin biosynthetic process, regulation of chromosome organization, regulation of flower development, regulation of pollen tube growth, regulation of response to water deprivation, regulation of seed germination, reproduction, reproductive structure development, response to cold, response to freezing, response to gamma radiation, response to heat, response to high light intensity, response to light stimulus, response to salt stress, response to starvation, seed coat development, seed development, seed dormancy process, seed germination, seed maturation, seedling development, sepal formation, spindle assembly, stamen development, starch biosynthetic process, starch catabolic process, starch metabolic process, synapsis, vegetative phase change, vegetative to reproductive phase transition of meristem, vernalization response.

List of gene ontology terms selected for cyto-nuclear interaction related genes from 'cellular components' keywords: chloroplast, chloroplast envelope, chloroplast inner membrane, chloroplast membrane, chloroplast outer membrane, chloroplast starch grain, chloroplast stroma, chloroplast thylakoid, chloroplast thylakoid membrane, chromoplast, DNA topoisomerase complex (ATP-hydrolyzing), DNA-directed RNA polymerase II, holoenzyme, F-actin capping protein complex, GPI-anchor transamidase complex, integral component of chloroplast outer membrane, integral component of mitochondrial outer membrane, integral component of nuclear inner membrane, integral component of

thylakoid membrane, katanin complex, MCM complex, mediator complex, mitochondrial inner membrane, mitochondrial matrix, mitochondrial outer membrane translocase complex, mitochondrial proton-transporting ATP synthase complex, coupling factor F(o), mitochondrion, outer membrane, photosystem I antenna complex, photosystem II, photosystem II oxygen evolving complex, plastid, plastid envelope, plastid inner membrane, plastid outer membrane, plastoglobule, proton-transporting ATP synthase complex, catalytic core F(1), stromule, thylakoid, thylakoid lumen, thylakoid membrane.

3.7.3 Supplementary tables

Table 3.1 – Summary table of pre- and post-trimming, including GC content, read length and count of paired-end (PE) and single reads for the latter, and name of each library.

	H. hispanica	***H. non-scripta***	***H. paivae***	***H. hispanica***
Library name	SWA1	SWA2	SWA3	SWA4
Sampling ID	BB-339	BB-411	BB-130	BB-356
Sequencing technology	Illumina PE 2x 100bp	Illumina PE 2x 100bp	Illumina PE 2x 100bp	Roche 454 GS FLX
Raw data				
No. of reads	115,530,874	101,782,090	99,171,658	1,347,397
Length	10-100	10-100	10-100	20-1157
% GC	45	45	45	43
Trimmed reads				
No. of R1 reads	31,127,005	33,086,118	29,331,396	
Length	25-100	30-100	25-100	
GC%	46	46	46	
No. of R2 reads	31,127,005	33,086,118	29,331,396	
Length	25-100	25-100	25-100	
GC%	45	45	45	
No. of broken pairs	22,782,611	16,044,763	18,263,097	
Length	25-100	25-100	25-100	
GC%	43	41	41	
Total post-processed reads (PE + broken pairs)				
No. of reads	85,036,621	82,216,999	76,925,889	953,939
Length	25-100	25-100	25-100	86-686
Insert size PE ± SD	202 ± 68	240 ± 69	228 ± 65	
% GC	43-46	41-46	41-46	44

Table 3.2 – Summary table of assembled contigs by library, assemblers and their kmer sizes.

Library	Assembler	kmer	Contigs	Sum length	N50	Min	Length Max	Median	Mean	Sd
SWA4	Newbler	–	21,197	18,994,328	1,144	86	5,818	666	896	573
SWA4	SOAPdenovo-Trans	31	43,410	14,295,190	422	100	3,280	239	329	278
SWA4	SOAPdenovo-Trans	127	11,457	6,273,762	536	128	2,566	471	548	249
SWA4	SOAPdenovo-Trans	63	39,194	13,457,564	414	100	3,201	272	343	262
SWA1	SOAPdenovo-Trans	25	190,980	55,451,836	403	100	27,082	163	290	397
SWA1	SOAPdenovo-Trans	31	175,999	63,288,008	754	100	15,019	159	360	500
SWA1	SOAPdenovo-Trans	75	54,267	25,861,582	683	100	4,661	292	477	428
SWA2	SOAPdenovo-Trans	25	169,074	56,615,855	599	100	40,111	172	335	448
SWA2	SOAPdenovo-Trans	31	179,678	60,382,407	628	100	22,853	170	336	443
SWA2	SOAPdenovo-Trans	75	52,135	36,606,622	1,221	100	12,011	435	702	648
SWA3	SOAPdenovo-Trans	75	76,130	37,136,994	813	100	9,387	269	488	478
SWA3	SOAPdenovo-Trans	25	150,152	54,983,724	696	100	21,310	181	366	459
SWA3	SOAPdenovo-Trans	31	157,063	58,918,598	747	100	18,617	181	375	472
SWA1	Trinity	25	103,067	81,193,492	1,247	201	12,718	502	788	695
SWA2	Trinity	25	109,800	99,541,401	1,354	201	16,392	665	907	743
SWA3	Trinity	25	94,989	87,069,136	1,341	201	9,577	698	917	708
SWA4	Trinity	25	33,368	22,275,172	823	201	5,785	508	668	469
SWA1	Velvet/Oases	31	98,150	98,859,110	1,517	164	21,369	750	1,007	824
SWA2	Velvet/Oases	31	111,739	113,369,705	1,479	200	19,941	789	1,015	790
SWA3	Velvet/Oases	31	101,988	108,237,313	1,509	200	12,429	861	1,061	779

Table 3.3 – Summary table of redundancy removal and alignment rates.

	H. hispanica	***H. non-scripta***	***H. paivae***	***H. hispanica***
Library name	SWA1	SWA2	SWA3	SWA4
Sampling ID	BB-339	BB-411	BB-130	BB-356
Sequencing technology	Illumina PE 2x 100bp	Illumina PE 2x 100bp	Illumina PE 2x 100bp	Roche 454 GS FLX
Evidential gene tr2aacds pipeline				
Raw contigs	622,463	622,426	580,322	123,310
CDS (bestORF)	639,724	644,243	599,647	125,963
Primary transcripts	33,761	31,909	31,917	15,940
Alternative transcripts	29,344	28,104	25,273	7,943
Trimmed PE reads aligned to all their primary transcripts				
% aligned reads	66.27	64.07	61.46	32.56
% aln = 1 x	44.91	45.15	42.69	25.37
% aln >1 x	21.36	18.92	18.77	7.19
% proper PE alignment	69.36	70.41	68.69	–
Effective mean cov + SD	33.20 ± 39.07	38.54 ± 44.80	37.56 ± 40.14	5.39 ± 9.75
Median/max coverage	15.34/183.54	18.28/193.03	21.65/191.29	2.83/201.65
N loci ≥ 10X eff. cov	19,832	20,241	22,273	1554
Trimmed reads aligned to the top 1000 longest transcripts (statistics of the contig lengths)				
Mean	943	921	869	560
Median	893	877	830	528
Min	767	756	720	452
Max	2189	2375	1750	1086

Table 3.4 – Effect of filtering steps and final number of 236 genes with 376 target SNPs in a total of 300 different amplicon regions. * selected single-copy genes with one or few exons in *Phoenix dactylifera*.

	Nuclear		**Chloroplast**		**Mitochondrion**	
	Genes	**SNPs**	**Genes**	**SNPs**	**Genes**	**SNPs**
Transcripts mapped to *Musa acuminata/Phoenix dactylifera* with >80 % coverage of length and shared between three bluebells	1,047	20,939	86	120	40	13
Fixed polymorphisms between *H. hispanica* and *H. non-scripta*	730	3,331	39	96	9	10
SNP within exon boundaries of Musa acuminata	727	3,315	10*	10	6*	6
Regions provide 80 bp flanking region	275	581	-	-	-	-
Positive strand and genes filtered for GO annotations	232	447	10	10	6	6
Sites for which primer oligos were successful	222	374	10	15	5	5
Final set	**221**	**361**	**10**	**10**	**5**	**5**

Table 3.5 – Chloroplastic genes discovered from *de novo* assembly and RNAseq alignment to *Phoenix dactylifera* (Khan et al., 2011). A – genes, for which amplicons were designed (x). R – genes, which successfully amplified and were polymorphic (x).

Reference name P. dactylifera|bluebell_contig	Gene	Gene info	A	R
NC_013991.2_cds_YP_003540911.1_3|SWA2_509241	psbA	Photosystem protein genes	x	x
NC_013991.2_cds_YP_003540916.1_10|SWA2_466094	atpA	ATP synthase subunit genes	x	x
NC_013991.2_cds_YP_003540921.1_15|SWA2_65567	rpoC2	RNA polymerase genes	x	x
NC_013991.2_cds_YP_003540938.1_32|SWA2_6166	atpB	ATP synthase subunit genes	x	x
NC_013991.2_cds_YP_003540939.1_33|SWA2_98197utrorf	rbcL	Rubisco subunit gene	x	x
NC_013991.2_cds_YP_003540940.1_34|SWA2_476247	accD	Acetyl-CoA-carboxylase subunit gene	x	x
NC_013991.2_cds_YP_003540943.2_37|SWA2_458496utrorf	cemA	Envelope membrane protein gene	x	x
NC_013991.2_cds_YP_003540971.1_68|SWA2_260076_rpl22	rpl22	Ribosomal protein genes	x	x
NC_013991.2_cds_YP_003540920.1_14|SWA2_65566	rps2	Ribosomal protein genes	x	0
NC_013991.2_cds_YP_003540967.1_63|SWA2	rps8	Ribosomal protein genes	x	0
NC_013991.2_cds_YP_003540910.1_2|SWA2	rps12	Ribosomal protein genes 3 different ones	0	0
NC_013991.2_cds_YP_003540912.1_4|SWA2_46035	matK	Maturase gene	0	0
NC_013991.2_cds_YP_003540913.1_6|SWA2	rps16	Ribosomal protein genes *one intron	0	0
NC_013991.2_cds_YP_003540914.1_7|SWA2	psbK	Photosystem protein genes	0	0
NC_013991.2_cds_YP_003540917.1_11|SWA2	atpF	ATP synthase subunit genes	0	0
NC_013991.2_cds_YP_003540918.1_12|SWA2	atpH	ATP synthase subunit genes	0	0
NC_013991.2_cds_YP_003540919.1_13|SWA2_65565utrorf	atpI	ATP synthase subunit genes	0	0
NC_013991.2_cds_YP_003540922.1_16|SWA2_458538utrorf	rpoC1	RNA polymerase genes *one intron	0	0
NC_013991.2_cds_YP_003540923.1_17|SWA2_65567utrorf	rpoB	RNA polymerase genes	0	0
NC_013991.2_cds_YP_003540926.1_20|SWA2_457676	psbD	Photosystem protein genes	0	0
NC_013991.2_cds_YP_003540927.1_21|SWA2_100344	psbC	Photosystem protein genes	0	0
NC_013991.2_cds_YP_003540928.1_22|SWA2	psbZ	Photosystem protein genes	0	0
NC_013991.2_cds_YP_003540929.2_23|SWA2	rps14	Ribosomal protein genes	0	0
NC_013991.2_cds_YP_003540930.1_24|SWA2_51927utrorf	psaB	Photosystem protein genes	0	0
NC_013991.2_cds_YP_003540931.1_25|SWA2_519268	psaA	Photosystem protein genes	0	0
NC_013991.2_cds_YP_003540932.1_26|SWA2	ycf3	Genes of unknown function ** 2 introns	0	0
NC_013991.2_cds_YP_003540933.1_27|SWA2_58783utrorf	rps4	Ribosomal protein genes	0	0
NC_013991.2_cds_YP_003540934.1_28|SWA2	ndhJ	NADH subunits genes	0	0
NC_013991.2_cds_YP_003540935.1_29|SWA2_472285	ndhK	NADH subunits genes	0	0
NC_013991.2_cds_YP_003540936.1_30|SWA2	ndhC	NADH subunits genes	0	0
NC_013991.2_cds_YP_003540937.1_31|SWA2	atpE	ATP synthase subunit genes	0	0
NC_013991.2_cds_YP_003540942.1_36|SWA2_556910	ycf4	Genes of unknown function	0	0
NC_013991.2_cds_YP_003540944.1_38|SWA2_458496	petA	Cytochrome-related gene	0	0
NC_013991.2_cds_YP_003540946.1_40|SWA2	psbL	Photosystem protein genes	0	0
NC_013991.2_cds_YP_003540948.1_42|SWA2	psbE	Photosystem protein genes	0	0
NC_013991.2_cds_YP_003540951.1_45|SWA2	psaJ	Photosystem protein genes	0	0
NC_013991.2_cds_YP_003540952.1_46|SWA2	rpl33	Ribosomal protein genes	0	0
NC_013991.2_cds_YP_003540953.1_47|SWA2_444220	rps18	Ribosomal protein genes	0	0
NC_013991.2_cds_YP_003540954.1_48|SWA1_522004	rpl20	Ribosomal protein genes *one intron	0	0
NC_013991.2_cds_YP_003540955.2_49|SWA2	rps12	Ribosomal protein genes 3 different ones	0	0
NC_013991.2_cds_YP_003540956.1_50|SWA2_512055	clpP	Proteosome like protease gene **two introns	0	0
NC_013991.2_cds_YP_003540957.1_51|SWA2_71517	psbB	Photosystem protein genes	0	0
NC_013991.2_cds_YP_003540959.1_53|SWA2	psbN	Photosystem protein genes	0	0
NC_013991.2_cds_YP_003540960.1_54|SWA2	psbH	Photosystem protein genes	0	0
NC_013991.2_cds_YP_003540961.1_56|SWA2	petB	Cytochrome-related gene *one intron	0	0
NC_013991.2_cds_YP_003540962.1_58|SWA2	petD	Cytochrome-related gene	0	0
NC_013991.2_cds_YP_003540963.1_59|SWA2_458383	rpoA	RNA polymerase genes	0	0
NC_013991.2_cds_YP_003540964.1_60|SWA2	rps11	Ribosomal protein genes	0	0
NC_013991.2_cds_YP_003540965.1_61|SWA2	rpl36	Ribosomal protein genes	0	0
NC_013991.2_cds_YP_003540966.1_62|SWA2	infA	Translation initiation factor gene	0	0
NC_013991.2_cds_YP_003540968.1_64|SWA2	rpl14	Ribosomal protein genes	0	0
NC_013991.2_cds_YP_003540969.1_66|SWA2	rpl16	Ribosomal protein genes *one intron	0	0
NC_013991.2_cds_YP_003540970.1_67|SWA2_273607_rps3	rps3	Ribosomal protein genes	0	0
NC_013991.2_cds_YP_003540972.1_69|SWA2	rps19	Ribosomal protein genes x2	0	0
NC_013991.2_cds_YP_003540973.2_70|SWA2_71504	rpl2	Ribosomal protein genes *one intron x2	0	0
NC_013991.2_cds_YP_003540974.1_71|SWA2	rpl23	Ribosomal protein genes x2	0	0
NC_013991.2_cds_YP_003540975.1_72|SWA2	ycf2	Genes of unknown function x2	0	0
NC_013991.2_cds_YP_003540976.1_73|SWA2	ndhB	NADH subunits genes *one intron x2 in IR	0	0
NC_013991.2_cds_YP_003540977.1_74|SWA2	rps7	Ribosomal protein genes x2	0	0
NC_013991.2_cds_YP_003540978.1_75|SWA2	ycf68	Genes of unknown function *one intron x2	0	0
NC_013991.2_cds_YP_003540979.1_76|SWA2_71725utrorf	ycf1	Genes of unknown function	0	0
NC_013991.2_cds_YP_003540980.1_77|SWA1_323734	ndhF	NADH subunits genes	0	0
NC_013991.2_cds_YP_003540982.1_79|SWA2_511600	ccsA	Cytochrome-related gene	0	0
NC_013991.2_cds_YP_003540983.1_80|SWA2_49235	ndhD	NADH subunits genes	0	0
NC_013991.2_cds_YP_003540984.1_81|SWA2	psaC	Photosystem protein genes	0	0
NC_013991.2_cds_YP_003540985.1_82|SWA2	ndhE	NADH subunits genes	0	0
NC_013991.2_cds_YP_003540986.1_83|SWA2_512222	ndhG	NADH subunits genes	0	0
NC_013991.2_cds_YP_003540987.1_84|SWA2	ndhI	NADH subunits genes	0	0
NC_013991.2_cds_YP_003540988.1_85|SWA2_512219_cns_ndhA	ndhA	NADH subunits genes *one intron	0	0
NC_013991.2_cds_YP_003540989.1_86|SWA2_286651	ndhH	NADH subunits genes	0	0
NC_013991.2_cds_YP_003540991.1_88|SWA2_476818	ycf1	Genes of unknown function	0	0
NC_013991.2_cds_YP_003540992.1_89|SWA2	ycf68	Genes of unknown function *one intron x2	0	0
NC_013991.2_cds_YP_003540993.1_90|SWA2	rps7	Ribosomal protein genes x2	0	0
NC_013991.2_cds_YP_003540994.1_91|SWA2_458483_rps12	ndhB	NADH subunits genes *one intron x2 in IR	0	0
NC_013991.2_cds_YP_003540994.1_91|SWA2_71510utrorf_ndhB	ndhB	NADH subunits genes *one intron x2 in IR	0	0
NC_013991.2_cds_YP_003540995.1_92|SWA2_279270_cns	ycf2	Genes of unknown function	0	0
NC_013991.2_cds_YP_003540996.1_93|SWA2	rpl23	Ribosomal protein genes x2	0	0

NC_013991.2_cds_YP_003540997.1_94|SWA2_71504	rpl2	Ribosomal protein genes *one intron x2	0	0
NC_013991.2_cds_YP_003540998.1_95|SWA2	rps19	Ribosomal protein genes x2	0	0
NC_013991.2_cds_YP_003778185.1_1|SWA2	rps12	Ribosomal protein genes 3 different ones	0	0
NC_013991.2_cds_YP_003778186.1_5|SWA2	rps16	Ribosomal protein genes *one intron	0	0
NC_013991.2_cds_YP_003778188.1_55|SWA2_275442	petB	Cytochrome-related gene *one intron	0	0
NC_013991.2_cds_YP_003778189.1_57|SWA2	petD	Cytochrome-related gene	0	0
NC_013991.2_cds_YP_003778190.1_65|SWA1_525852	rpl16	Ribosomal protein genes *one intron	0	0

Table 3.6 – Mitochondrial genes discovered from *de novo* assembly and RNAseq alignment to *Phoenix dactylifera* (Fang et al., 2012). A – genes, for which amplicons were designed (x). R – genes, which successfully amplified and were polymorphic (x).

Reference name *P. dactylifera*	**Gene**	**Gene name**	**Group of genes**	**A**	**R**
NC_016740.1_cds_YP_005090359.1_2	atp6	ATP synthase F0 subunit 6	Complex V *mt origin	x	x
NC_016740.1_cds_YP_005090365.1_8	cox1	cytochrome c oxidase subunit 1	*mt origin	x	x
NC_016740.1_cds_YP_005090369.1_12	nad1	NADH dehydrogenase subunit 1	Complex I *mt origin	x	x
NC_016740.1_cds_YP_005090375.1_18	atp4	ATP synthase F0 subunit 4	Complex V *mt origin	x	x
NC_016740.1_cds_YP_005090389.1_32	nad5	NADH dehydrogenase subunit 5	Complex I *mt origin	x	x
NC_016740.1_cds_YP_005090358.1_1	nad2	NADH dehydrogenase subunit 2	Complex I *mt origin	0	0
NC_016740.1_cds_YP_005090360.1_3	rpl5	ribosomal protein L5	Ribosome large subunit	0	0
NC_016740.1_cds_YP_005090361.1_4	rps14	ribosomal protein S14	Ribosome small subunit	0	0
NC_016740.1_cds_YP_005090362.1_5	cob	apocytochrome b	Complex III *mt origin	0	0
NC_016740.1_cds_YP_005090364.1_7	rps7	ribosomal protein S7	Ribosome small subunit	0	0
NC_016740.1_cds_YP_005090366.1_9	nad7	NADH dehydrogenase subunit 7	Complex I *mt origin	0	0
NC_016740.1_cds_YP_005090367.1_10	orf186	orf186	Hypothetical genes	0	0
NC_016740.1_cds_YP_005090368.1_11	rps13	ribosomal protein S13	Ribosome small subunit	0	0
NC_016740.1_cds_YP_005090370.1_13	nad6	NADH dehydrogenase subunit 6	Complex I *mt origin	0	0
NC_016740.1_cds_YP_005090372.1_15	rps19	ribosomal protein S19	Ribosome small subunit	0	0
NC_016740.1_cds_YP_005090373.1_16	rps3	ribosomal protein S3	Ribosome small subunit	0	0
NC_016740.1_cds_YP_005090374.1_17	rpl16	ribosomal protein L16	Ribosome large subunit	0	0
NC_016740.1_cds_YP_005090376.1_19	nad4L	NADH dehydrogenase subunit 4L	Complex I *mt origin	0	0
NC_016740.1_cds_YP_005090377.1_20	ccmFc	cytochrome c biogenesis FC	*mt origin	0	0
NC_016740.1_cds_YP_005090378.1_21	atp1	ATP synthase F0 subunit 1	Complex V *mt origin	0	0
NC_016740.1_cds_YP_005090379.1_22	atp9	ATP synthase F0 subunit 9	Complex V *mt origin	0	0
NC_016740.1_cds_YP_005090380.1_23	rps1	ribosomal protein S1	Ribosome small subunit	0	0
NC_016740.1_cds_YP_005090381.1_24	nad4	NADH dehydrogenase subunit 4	Complex I *mt origin	0	0
NC_016740.1_cds_YP_005090382.1_25	cox3	cytochrome c oxidase subunit 3	*mt origin	0	0
NC_016740.1_cds_YP_005090383.1_26	rps2	ribosomal protein S2	Ribosome small subunit	0	0
NC_016740.1_cds_YP_005090384.1_27	rps4	ribosomal protein S4	Ribosome small subunit	0	0
NC_016740.1_cds_YP_005090385.1_28	rps11	ribosomal protein S11	Ribosome small subunit	0	0
NC_016740.1_cds_YP_005090386.1_29	ccmFn	cytochrome c biogenesis Fn	*mt origin	0	0
NC_016740.1_cds_YP_005090387.1_30	mttB	MttB	SecY-independent transporter	0	0
NC_016740.1_cds_YP_005090388.1_31	cox2	cytochrome c oxidase subunit 2	*mt origin	0	0
NC_016740.1_cds_YP_005090390.1_33	ccmB	cytochrome c biogenesis B	*mt origin	0	0
NC_016740.1_cds_YP_005090391.1_34	orf100	orf100	Hypothetical genes	0	0
NC_016740.1_cds_YP_005090392.1_35	matR	maturase	intron maturase *mt origin	0	0
NC_016740.1_cds_YP_005090393.1_36	orf192	orf192	Hypothetical genes	0	0
NC_016740.1_cds_YP_005090394.1_37	atp8	ATP synthase F0 subunit 8	Complex V *mt origin	0	0
NC_016740.1_cds_YP_005090395.1_38	orf142	orf142	Hypothetical genes	0	0
NC_016740.1_cds_YP_005090396.1_39	nad9	NADH dehydrogenase subunit 9	Complex I *mt origin	0	0
NC_016740.1_cds_YP_005090397.1_40	ccmC	cytochrome c biogenesis C	*mt origin	0	0
NC_016740.1_cds_YP_005090398.1_41	nad3	NADH dehydrogenase subunit 3	Complex I *mt origin	0	0
NC_016740.1_cds_YP_005090399.1_42	rps12	ribosomal protein S12	Ribosome small subunit	0	0

Table 3.7 – Subset of 221 nuclear genes discovered from *de novo* assembly and comparison with *Musa acuminata*. A – genes, for which amplicons were designed (x). R – genes, which successfully amplified and were polymorphic (x).

Reference name *M. acuminata*	**Gene info**	**A**	**R**
GSMUA_Achr4G01020	Elongation factor 2	X	0
GSMUA_Achr4G18340	Putative UPF0420 protein C16orf58 homolog	X	0
GSMUA_Achr4G29520	Alcohol dehydrogenase-like 3	X	0
GSMUA_Achr5G20990	S-adenosylmethionine synthase	X	0
GSMUA_Achr6G30620	Putative Aspartic proteinase Asp1	X	0
GSMUA_Achr9G01230	Pentatricopeptide repeat-containing protein At2g13420, mitochondrial	X	0
GSMUA_Achr9G03370	Putative Ribosomal RNA large subunit methyltransferase E	X	0
GSMUA_Achr10G00370	Magnesium-protoporphyrin IX monomethyl ester [oxidative] cyclase, chloroplastic	X	X
GSMUA_Achr10G08820	exostosin, putative, expressed	X	X
GSMUA_Achr10G11820	Putative [Protein-PII] uridylyltransferase	X	X
GSMUA_Achr10G13340	Protein ALUMINUM SENSITIVE 3	X	X
GSMUA_Achr10G14120	Whole genome shotgun sequence of line PN40024	X	X
GSMUA_Achr10G14600	DEAD-box ATP-dependent RNA helicase 38	X	X
GSMUA_Achr10G14700	Whole genome shotgun sequence of line PN40024	X	X
GSMUA_Achr10G14800	Pentatricopeptide repeat-containing protein At5g25630	X	X
GSMUA_Achr10G17440	ubiquitin carboxyl-terminal hydrolase, family 1, putative expressed	X	X
GSMUA_Achr10G17950	ATMAP70 protein, putative, expressed	X	X
GSMUA_Achr10G19980	Putative 66 kDa stress protein	X	X
GSMUA_Achr10G20800	methyltransferase domain containing protein, expressed	X	X
GSMUA_Achr10G24620	Putative Probable nitronate monooxygenase	X	X
GSMUA_Achr10G26390	Probable chlorophyll(ide) b reductase NYC1, chloroplastic	X	X
GSMUA_Achr10G26470	peptidase C45, acyl-coenzyme A/6-aminopenicillanic acid acyl-transferase, putative expressed	X	X
GSMUA_Achr10G27360	sterol glucosyltransferase, putative, expressed	X	X
GSMUA_Achr10G27410	Putative expressed protein	X	X
GSMUA_Achr10G27650	Putative tRNA A64-2'-O-ribosylphosphate transferase	X	X
GSMUA_Achr10G28780	zinc ion binding protein, putative, expressed	X	X
GSMUA_Achr10G29540	Formate–tetrahydrofolate ligase	X	X
GSMUA_Achr10G31020	Whole genome shotgun sequence of line PN40024	X	X
GSMUA_Achr11G05060	Putative Pentatricopeptide repeat-containing protein At4g01990 , mitochondrial	X	X
GSMUA_Achr11G05770	Putative Serine/threonine-protein kinase-like protein ACR4	X	X
GSMUA_Achr11G08150	Putative Lysine-8-amino-7-oxononanoate aminotransferase	X	X
GSMUA_Achr11G08830	Biotin synthase	X	X
GSMUA_Achr11G10660	NFD4, putative, expressed	X	X
GSMUA_Achr11G12990	DUF630/DUF632 domains containing protein, putative, expressed	X	X
GSMUA_Achr11G15240	Putative Probable glucan endo-1,3-beta-glucosidase A6	X	X
GSMUA_Achr11G15710	Putative Spermatogenesis-associated protein 20	X	X
GSMUA_Achr11G16300	Putative expressed protein	X	X
GSMUA_Achr11G16420	Putative Pentatricopeptide repeat-containing protein At1g08610	X	X
GSMUA_Achr11G18720	Probable prenylcysteine oxidase	X	X
GSMUA_Achr11G22050	RuBisCO large subunit-binding protein subunit alpha, chloroplastic	X	X
GSMUA_Achr11G23470	Putative Anaphase-promoting complex subunit cdc20	X	X
GSMUA_Achr11G23880	Putative Pentatricopeptide repeat-containing protein At2g37320	X	X
GSMUA_Achr1G00740	Putative Crooked neck-like protein 1	X	X
GSMUA_Achr1G01980	Cytokinin dehydrogenase 3	X	X
GSMUA_Achr1G02260	Putative Glutathione S-transferase ERD13	X	X
GSMUA_Achr1G04970	Mannan endo-1,4-beta-mannosidase 2	X	X
GSMUA_Achr1G05790	Alcohol dehydrogenase class-3	X	X
GSMUA_Achr1G07730	Putative 1-acylglycerophosphocholine O-acyltransferase 1	X	X
GSMUA_Achr1G08300	transporter-related, putative, expressed	X	X
GSMUA_Achr1G09520	F-box/kelch-repeat protein At5g15710	X	X
GSMUA_Achr1G13350	expressed protein	X	X
GSMUA_Achr1G13560	Pentatricopeptide repeat-containing protein At1g26460, mitochondrial	X	X
GSMUA_Achr1G14120	BI1-like protein	X	X
GSMUA_Achr1G15770	Putative Probable sodium-coupled neutral amino acid transporter 6	X	X
GSMUA_Achr1G18060	Putative Rhamnogalacturonate lyase	X	X
GSMUA_Achr1G18310	calmodulin binding protein, putative, expressed	X	X
GSMUA_Achr1G20950	Alpha-xylosidase	X	X
GSMUA_Achr1G23020	Uncharacterized PKHD-type hydroxylase At1g22950	X	X
GSMUA_Achr1G26020	Pyrophosphate-energized vacuolar membrane proton pump	X	X
GSMUA_Achr1G27940	ternary complex factor MIP1, putative, expressed	X	X
GSMUA_Achr2G02770	Naringenin,2-oxoglutarate 3-dioxygenase (Fragment)	X	X
GSMUA_Achr2G07310	remorin C-terminal domain containing protein, putative, expressed	X	X
GSMUA_Achr2G08260	expressed protein	X	X
GSMUA_Achr2G08450	Protein HOTHEAD	X	X
GSMUA_Achr2G09000	Putative SNF1-related protein kinase regulatory subunit gamma 1	X	X
GSMUA_Achr2G09070	GTP binding protein, putative, expressed	X	X
GSMUA_Achr2G09160	DEAD-box ATP-dependent RNA helicase 35	X	X
GSMUA_Achr2G09290	Putative expressed protein	X	X
GSMUA_Achr2G09300	Chlorophyll a-b binding protein 3C, chloroplastic	X	X
GSMUA_Achr2G13970	ATP synthase gamma chain 1, chloroplastic	X	X
GSMUA_Achr2G15280	Whole genome shotgun sequence of line PN40024	X	X
GSMUA_Achr2G15610	Putative Serine/threonine-protein kinase HT1	X	X

Table 3.7 continued

Reference name *M. acuminata*	**Gene info**	**A**	**R**
GSMUA_Achr2G19670	Putative glycosyl transferase, group 1 domain containing protein, expressed	X	X
GSMUA_Achr2G22450	Phosphatidylinositol-4-phosphate 5-kinase 1	X	X
GSMUA_Achr3G02410	Scarecrow-like protein 1	X	X
GSMUA_Achr3G03330	Molybdopterin biosynthesis protein CNX3	X	X
GSMUA_Achr3G03420	Putative Glyoxylate reductase	X	X
GSMUA_Achr3G04190	Cytochrome c oxidase assembly protein COX15 homolog	X	X
GSMUA_Achr3G05380	Aspartyl-tRNA synthetase, cytoplasmic	X	X
GSMUA_Achr3G09370	Putative Subtilisin-like protease	X	X
GSMUA_Achr3G09850	Putative Predicted protein	X	X
GSMUA_Achr3G10020	Predicted protein	X	X
GSMUA_Achr3G11810	Ras-related protein RGP1	X	X
GSMUA_Achr3G14550	Putative Lipase member N	X	X
GSMUA_Achr3G14740	Putative DUF246 domain-containing protein At1g04910	X	X
GSMUA_Achr3G16150	6-phosphogluconate dehydrogenase, decarboxylating	X	X
GSMUA_Achr3G16560	T-complex protein 11, putative, expressed	X	X
GSMUA_Achr3G17610	Glycerol-3-phosphate acyltransferase 6	X	X
GSMUA_Achr3G19470	expressed protein	X	X
GSMUA_Achr3G21550	cyclin-related protein, putative, expressed	X	X
GSMUA_Achr3G23800	stress regulated protein, putative, expressed	X	X
GSMUA_Achr3G24620	Putative Probable glutamyl-tRNA synthetase, cytoplasmic	X	X
GSMUA_Achr3G25460	Probable RNA helicase SDE3	X	X
GSMUA_Achr3G25670	Putative Uncharacterized membrane protein At3g27390	X	X
GSMUA_Achr3G27080	Putative Intracellular protease 1	X	X
GSMUA_Achr3G28000	Probable xyloglucan glycosyltransferase 5	X	X
GSMUA_Achr3G29860	Putative Zinc finger CCCH domain-containing protein 18	X	X
GSMUA_Achr4G01380	Malate dehydrogenase, glyoxysomal	X	X
GSMUA_Achr4G03140	Putative Receptor protein kinase CLAVATA1	X	X
GSMUA_Achr4G03240	Putative Uncharacterized mitochondrial carrier YMR166C	X	X
GSMUA_Achr4G05330	Vacuolar protein sorting-associated protein 41 homolog	X	X
GSMUA_Achr4G11890	calcium-binding mitochondrial protein anon-60Da, putative, expressed	X	X
GSMUA_Achr4G11940	Luminal-binding protein 4	X	X
GSMUA_Achr4G14360	Putative Replication protein A 70 kDa DNA-binding subunit	X	X
GSMUA_Achr4G20810	ubiquitin carboxyl-terminal hydrolase, family 1, putative, expressed	X	X
GSMUA_Achr4G21470	5-methyltetrahydropteroyltriglutamate–homocysteine methyltransferase	X	X
GSMUA_Achr4G23090	Xylulose kinase	X	X
GSMUA_Achr4G26000	Putative Mitochondrial carnitine/acylcarnitine carrier protein CACL	X	X
GSMUA_Achr4G28160	Bifunctional aspartokinase/homoserine dehydrogenase 1, chloroplastic	X	X
GSMUA_Achr5G01650	Heparanase-like protein 1	X	X
GSMUA_Achr5G02150	pleckstrin homology domain-containing protein, putative, expressed	X	X
GSMUA_Achr5G03070	GTP binding protein, putative, expressed	X	X
GSMUA_Achr5G06050	Probable cellulose synthase A catalytic subunit 1 [UDP-forming]	X	X
GSMUA_Achr5G06790	expressed protein	X	X
GSMUA_Achr5G07900	UBA/TS-N domain containing protein, expressed	X	X
GSMUA_Achr5G14950	nodulin, putative, expressed	X	X
GSMUA_Achr5G15010	Putative Peptide transporter PTR2	X	X
GSMUA_Achr5G16020	Auxin-induced protein PCNT115	X	X
GSMUA_Achr5G18070	Alpha,alpha-trehalose-phosphate synthase [UDP-forming] 5	X	X
GSMUA_Achr5G18110	Probable inosine-5'-monophosphate dehydrogenase	X	X
GSMUA_Achr5G29660	Putative Probable serine/threonine-protein kinase NAK	X	X
GSMUA_Achr6G01600	asp/Glu racemase, putative, expressed	X	X
GSMUA_Achr6G03670	Putative KH domain-containing protein At4g18375	X	X
GSMUA_Achr6G03760	OsSBeL1 - Putative Serine Beta-Lactamase homologue, expressed	X	X
GSMUA_Achr6G04590	Putative Subtilisin-like protease	X	X
GSMUA_Achr6G06370	Sphingosine-1-phosphate lyase	X	X
GSMUA_Achr6G06520	Serine hydroxymethyltransferase 1	X	X
GSMUA_Achr6G06770	T-complex protein 1 subunit epsilon	X	X
GSMUA_Achr6G08110	Obtusifoliol 14-alpha demethylase	X	X
GSMUA_Achr6G08280	Putative Staphylococcal nuclease domain-containing protein 1	X	X
GSMUA_Achr6G09110	Probable monodehydroascorbate reductase, cytoplasmic isoform 2	X	X
GSMUA_Achr6G10570	Putative MYST-like histone acetyltransferase 1	X	X
GSMUA_Achr6G14730	Aquaporin NIP1-1	X	X
GSMUA_Achr6G16650	Diacylglycerol kinase 1	X	X
GSMUA_Achr6G18110	ubiquitin carboxyl-terminal hydrolase, family 1, putative	X	X
GSMUA_Achr6G18680	protein kinase family protein, putative	X	X
GSMUA_Achr6G21640	calcium-binding EF hand family protein, putative, expressed	X	X
GSMUA_Achr6G21970	Putative Formimidoyltransferase-cyclodeaminase	X	X
GSMUA_Achr6G22680	Sugar transport protein 14	X	X
GSMUA_Achr6G23690	Putative Nodulation protein H	X	X
GSMUA_Achr6G24240	Potassium channel AKT2/3	X	X
GSMUA_Achr6G24920	C2 domain containing protein, putative, expressed	X	X
GSMUA_Achr6G25590	Chlorophyllide a oxygenase, chloroplastic	X	X
GSMUA_Achr6G27910	Nucleolar GTP-binding protein 2	X	X
GSMUA_Achr6G28030	Whole genome shotgun sequence of line PN40024	X	X
GSMUA_Achr6G28840	GDSL esterase/lipase At3g26430	X	X
GSMUA_Achr6G33840	Alpha-glucan phosphorylase, H isozyme	X	X
GSMUA_Achr6G35130	Probable galactinol–sucrose galactosyltransferase 6	X	X
GSMUA_Achr6G35510	Putative Pentatricopeptide repeat-containing protein At1g22960, mitochondrial	X	X
GSMUA_Achr7G01860	Probable galactinol–sucrose galactosyltransferase 2	X	X
GSMUA_Achr7G02370	Glutamate decarboxylase	X	X
GSMUA_Achr7G02450	TPR repeat-containing thioredoxin TDX	X	X

Table 3.7 continued

Reference name *M. acuminata*	**Gene info**	**A**	**R**
GSMUA_Achr7G04520	2-phosphoglycerate kinase-related, putative, expressed	X	X
GSMUA_Achr7G07160	Probable mannitol dehydrogenase	X	X
GSMUA_Achr7G09140	retrotransposon protein, putative, unclassified, expressed	X	X
GSMUA_Achr7G10900	Branched-chain-amino-acid aminotransferase-like protein 3, chloroplastic	X	X
GSMUA_Achr7G11390	F-box/LRR-repeat protein 15	X	X
GSMUA_Achr7G13110	Putative Glucan endo-1,3-beta-glucosidase 10	X	X
GSMUA_Achr7G15560	proteins of unknown function domain containing protein, expressed	X	X
GSMUA_Achr7G16120	co-chaperone GrpE protein, putative, expressed	X	X
GSMUA_Achr7G16560	early-responsive to dehydration protein-related, putative, expressed	X	X
GSMUA_Achr7G18920	Alpha-amylase isozyme 3D	X	X
GSMUA_Achr7G19630	pyridoxamine 5'-phosphate oxidase family protein, putative, expressed	X	X
GSMUA_Achr7G20310	tetratricopeptide repeat domain containing protein, expressed	X	X
GSMUA_Achr7G21610	Putative Protein TRANSPARENT TESTA 12	X	X
GSMUA_Achr7G22520	Putative Probable importin subunit beta-4	X	X
GSMUA_Achr7G22690	Serine carboxypeptidase-like 27	X	X
GSMUA_Achr7G23770	expressed protein	X	X
GSMUA_Achr7G24000	Probable protein phosphatase 2C 6	X	X
GSMUA_Achr7G24470	Ketol-acid reductoisomerase, chloroplastic	X	X
GSMUA_Achr8G00180	SUMO-activating enzyme subunit 2	X	X
GSMUA_Achr8G00600	Putative Uncharacterized protein At2g33490	X	X
GSMUA_Achr8G03060	regulator of chromosome condensation/beta-lactamase-inhibitor protein II, putative expressed	X	X
GSMUA_Achr8G04580	T-complex protein 1 subunit theta	X	X
GSMUA_Achr8G06320	Putative F-box protein At5g49610	X	X
GSMUA_Achr8G10290	hydrolase, alpha/beta fold family domain containing protein, expressed	X	X
GSMUA_Achr8G12790	Putative expressed protein	X	X
GSMUA_Achr8G14130	Putative Epoxide hydrolase 2	X	X
GSMUA_Achr8G14690	Oligopeptide transporter 4	X	X
GSMUA_Achr8G15270	Pentatricopeptide repeat-containing protein At1g18900	X	X
GSMUA_Achr8G16870	Putative DUF246 domain-containing protein At1g04910	X	X
GSMUA_Achr8G18690	Serine/threonine-protein kinase HT1	X	X
GSMUA_Achr8G19180	expressed protein	X	X
GSMUA_Achr8G19780	Putative Protein Mpv17	X	X
GSMUA_Achr8G19830	hAT dimerisation domain-containing protein, putative, expressed	X	X
GSMUA_Achr8G21050	glycosyltransferase family 43 protein, putative, expressed	X	X
GSMUA_Achr8G24820	Putative Alkylated DNA repair protein alkB homolog 8	X	X
GSMUA_Achr8G28550	Pentatricopeptide repeat-containing protein At3g48250, chloroplastic	X	X
GSMUA_Achr8G28770	Putative Transmembrane protein 136	X	X
GSMUA_Achr8G28940	F-box/LRR-repeat protein 12	X	X
GSMUA_Achr8G30730	Glutamate decarboxylase 1	X	X
GSMUA_Achr8G32660	glycosyl transferase, putative, expressed	X	X
GSMUA_Achr9G00320	DNA repair metallo-beta-lactamase, putative, expressed	X	X
GSMUA_Achr9G00720	Hypothetical protein	X	X
GSMUA_Achr9G01560	Ubiquitin carboxyl-terminal hydrolase 5	X	X
GSMUA_Achr9G05680	MAC/Perforin domain containing protein, putative, expressed	X	X
GSMUA_Achr9G06330	tesmin/TSO1-like CXC domain containing protein, expressed	X	X
GSMUA_Achr9G06890	Probable receptor-like protein kinase At2g42960	X	X
GSMUA_Achr9G08360	Amino acid permease 1	X	X
GSMUA_Achr9G08500	Putative [Protein-PII] uridylyltransferase	X	X
GSMUA_Achr9G09050	exostosin family domain containing protein, expressed	X	X
GSMUA_Achr9G09300	Putative Folic acid synthesis protein fol1	X	X
GSMUA_Achr9G10420	DEAD-box ATP-dependent RNA helicase 41	X	X
GSMUA_Achr9G10640	Auxin-induced protein PCNT115	X	X
GSMUA_Achr9G13690	Putative Protein VERNALIZATION INSENSITIVE 3	X	X
GSMUA_Achr9G13910	Putative Subtilisin-like protease	X	X
GSMUA_Achr9G20100	Putative Peptide transporter PTR1	X	X
GSMUA_Achr9G21140	BBF1 - 2 Bric-a-Brac, Tramtrack, Broad Complex BTB domains with a F5/8 type C discoidin domain, expressed	X	X
GSMUA_Achr9G24290	Alpha-1,4-galacturonosyltransferase 1	X	X
GSMUA_Achr9G24550	Sugar carrier protein C	X	X
GSMUA_Achr9G27270	Probable mannitol dehydrogenase	X	X
GSMUA_Achr9G27580	DUF1680 domain containing protein, putative, expressed	X	X
GSMUA_AchrUn_randomG00330	Probable exocyst complex component 6	X	X
GSMUA_AchrUn_randomG07570	Auxin-induced protein 5NG4	X	X
GSMUA_AchrUn_randomG08370	Histidine kinase 3	X	X
GSMUA_AchrUn_randomG09460	Probable cellulose synthase A catalytic subunit 2 [UDP-forming]	X	X
GSMUA_AchrUn_randomG12680	Phytoene dehydrogenase, chloroplastic/chromoplastic	X	X
GSMUA_AchrUn_randomG17340	hydrolase, alpha/beta fold family domain containing protein, expressed	X	X
GSMUA_AchrUn_randomG18870	Omega-6 fatty acid desaturase, chloroplastic	X	X
GSMUA_AchrUn_randomG20670	Putative Probable E3 ubiquitin-protein ligase HERC1	X	X
GSMUA_AchrUn_randomG25560	Fe(2+) transport protein 1	X	X

Chapter 4

A natural bluebell hybrid zone in northern Spain

4.1 Introduction

Origin of hybrid zones. Hybrid zones have been defined as narrow regions where two distinguishable populations interbreed with varying degree of reproductive isolation (Arnold, 1997; Barton and Hewitt, 1985, 1989) and they arise as a consequence of primary intergradation or secondary contact (Curry, 2015; Endler, 1977). The distinction between primary or secondary contact is a question of the nature of the hybrid zone (Barton and Hewitt, 1985). Primary intergradation describes divergence of a continuous population into two entities due to parapatric speciation and the hybrid individuals are evidence of ancestral, and potentially ongoing gene flow (Harrison and Larson, 2016). Secondary contact describes a hybrid zone between clearly distinguishable entities that came into contact after a certain time of independent evolution, which caused divergence due to genetic drift, mutations and selection (Sedghifar et al., 2015). Most of the well-studied hybrid zones were found to be in secondary contact, e.g. house mouse *Mus musculus* x *Mus domesticus* in eastern Germany (Teeter et al., 2010), *Chorthippus parallelus* in the Pyrenees (Bella et al., 2007), and *Picea glauca* x *Picea engelmannii* in western North America (De La Torre et al., 2015). However, advances of molecular markers and simulation approaches lead to an increasing interest in parapatric speciation with gene flow (Doebeli and Dieckmann, 2003; Nosil, 2008; Papadopulos et al., 2011). Harrison and Larson (2016) suggested that more hybrid zones than assumed could represent primary intergradation (e.g. intertidial snails *Littorina*, Hollander et al., 2015).

European hybrid zones during Quaternary oscillations. Quaternary climatic oscillations have impacted the European biota (Comes and Kadereit, 2003; Hewitt, 1996, 2004). The last glacial cycle (about 135 ka B.P.) and the present interglacial warm period are best understood (Hewitt, 1996, and references therein). During the long glacial phases northern Europe was covered by ice, in addition, the mountain peaks were covered by glaciers, e.g. Sierra Nevada, Pyrennes, and Cantabrian Mountains (Hewitt, 1996; Rackham and Grove, 2001; Serrano et al., 2016; Zagwijn, 1992). Across Europe the plains were mostly tundra and cold steppe between the ice sheet at latitude 52°N and the European

mountain peaks (Hewitt, 1996). The species retreated or survived in so-called refugia of southern Europe, which provided suitable habitat within the Mediterranean basin (Brewer et al., 2002; Gómez and Lunt, 2006; Petit et al., 2003). Especially, the Iberian Peninsula has been suggested as a southern refugium to species (over the Quaternary oscillations O'Regan, 2008). Furthermore, dispersal patterns of uniparental genetic markers, frequent hybridisation, high endemism, and concordance of phylogeographic studies provide evidence for multiple refugia within the Iberian Peninsula. Multiple refugia within the southern Iberian refugium are possible because it is a geographical complex region with different biomes that provided refugia or sanctuaries throughout glacial cycles (Feliner, 2011; Gómez and Lunt, 2006; Recuero and Garcia-Paris, 2011; Taberlet et al., 1998). During inter-glacial phases, species re-colonised northern habitat, probably via routes described for several species (Hewitt, 1999; O'Regan, 2008). It has been shown for some trees, such as in genus *Quercus*, that since the last glacial maximum (LGM, 18 ka B.P., Bennett et al., 1991) species from southern Mediterranean refugia slowly recolonised central European mountains (late-glacial interstadial, 13-11 ka B.P.) and migrating increasingly to northern Europe after approximately 10 ka B.P. – the beginning of the Holocene (Brewer et al., 2002; Hewitt, 1999). Additionally, the surviving populations or species had diverged in allopatry and then formed secondary contact zones. Indeed, most hybrid zones are considered to have originated after the LGM and they are more likely to have persisted to this day. Nevertheless, hybrid zones with a parapatric origin can also be explained by the Pleistocene and Holocene glaciations influencing distribution range over smaller scales for example tree line shifts with mountain elevation (as opposed to a longitudinal N-S gradient from southern refugia (Hewitt, 1996). For instance, the Montes de León (northern part of the Galician-Duero Mountains) were suggested to promote parapatric speciation over Quaternary oscillations in the Iberian rock-lizards clade *Iberolacerta* based on strong genetic substructure in mtDNA of *I. monticola* (Remon et al., 2013).

Hybrid zones dynamics. Population dynamics in clinal hybrid zones (in contrast to patchy mosaic hybrid zones) are described as a balance between dispersal rate of the parental species into the centre, and the fitness of the hybrids (depending on environmental selection pressures or genetic interactions). Four models of hybrid zone models are suggested (Arnold, 1997): tension zone (Barton and Hewitt, 1985, 1989), bounded hybrid superiority (Moore, 1977), mosaic (Howard et al., 1993; Rand and Harrison, 1989), and evolutionary novelty (Arnold, 1997). Curry (2015) discussed these as a continuum dependent on geographic context and direction of selection on hybrids relative to the parental species. The models also focus on stable hybrid zones (Curry, 2015), although hybrid zones can shift their range – hybrid zone movement – due to asymmetric hybridisation, differential introgression and ecological changes, such as climate change (Buggs, 2007).

The fate of a hybrid zone can be the establishment of stable equilibrium reached by limited introgression at the hybrid zone margins, which maintains the pure parental populations (genetically as well as spatially). Alternatively, hybrid zones can collapse due to reinforced divergence, convergence with free inter-gradation of the genomes, or become replaced by a moving hybrid zone.

A neutral hybrid model assumes no directional selection and therefore a low genetic barrier separating the parents. As long as the hybrid zone is not trapped in low population density 'sinks', large amounts of introgression can be expected, leading to a uniform distribution of intermediate genotypes. The tension zone model assumes strong endogenous selection against hybrids and its maintenance is dependent on the dispersal of the parents into the hybrid zone, and is influenced by fitness dependant on the environment (Buggs, 2007). This type of zone presents substantial linkage disequilibrium and high heterozygote deficiency in the centre of the hybrid zone such that individuals have most alleles in multi-locus genotypes typical of one side of the zone, or most typical of the other - rather than being a 50:50 mixture of loci that are, apparently randomly segregating. Such distribution of genotypes (or phenotypes) is called bimodal. It also indicates a low amount of hybridisation or introgression (Gay et al., 2008). If the hybrid zone occurs along an environmental gradient, exogenous selection (i.e. selection imposed by the external environment) can influence the dynamics in two ways. If selection favours differential parental genotypes at either end of the environmental gradient (Endler, 1977), the genotype frequencies will resemble those of a tension zone. The strength of isolation would then depend on the steepness of environmental transition. Alternatively, enhanced fitness of the genotype of mixed ancestry along intermediate environments can also explain a steep clinal transition if there is a narrow band of habitat where the hybrids are favoured ('hybrid superiority model', Moore, 1977).

Reproductive barriers tested in hybrid zones. A hybrid zone can be used to explore reproductive barriers irrespectively of whether they were obtained in allo- or parapatry. One can distinguish barriers that are either effective before reproduction (pre-zygotic) or after reproduction (post-zygotic). Pre-zygotic barriers in plants inhibit the successful pollen transfer onto the stigma of a flower (including pollen growth tube). For instance, different perianth traits can affect pollinator preferences (Sheehan et al., 2012), e.g. a shift from bees to hummingbirds in *Mimulus lewisii* to *M. cardinalis* (Charlesworth and Charlesworth, 2000). Also, differences in flowering time (and selfing mating system) were suggested to promote reproductive isolation between three sympatric and cryptic lineages of *Juncus effusus* and *J. conglomeratus* (Michalski and Durka, 2015). Post-zygotic barriers in plants can occur at the genetic level and negatively influence the fitness of hybrids (Lafon-Placette et al., 2016). Lower fitness of hybrids can be caused by alleles of the parental species that are incompatible within the hybrid. Such Bateson-Dobzhansky-Muller incompatibilies were found in sunflower (*Helianthus*) backcrosses in combination with cyto-nuclear incompatibilities and local ecological preferences (Sambatti et al., 2008).

Cline analyses. Cline analyses can be used to quantify the strength of the genetic barrier and to trace introgression of genes from one species to the other, especially in hybrid zones. Allele frequencies per locus can be plotted along a spatial distance measure (geographic cline). The centre and width of geographic clines are particularly useful to determine genotypes associated with geography and environmental parameters. Moreover, the clines of molecular markers can be compared to other markers (cpDNA, morphology)

for concordance, which can reveal differential introgression and be used to identify loci that contribute to reproductive isolation (e.g. Stankowski et al., 2016).

Aims and objectives. Along the Southern foothills of the Cantabrian Mountains in Northern Spain, the British (*Hyacinthoides non-scripta* (L.) Chouard ex Rothm.) and Spanish (*Hyacinthoides hispanica* (Mill.) Rothm.) bluebell are naturally hybridising, which was first discovered by mixed plastid haplotypes within a locality (Grundmann et al., 2010). Conducting field collection, and re-sequencing of the developed genetic marker system for these samples, introgressive hybridisation with a potential influence of heterozygous advantage on the allele frequencies was investigated. Descriptive statistics of population genetics were used to quantify the inter-specific differentiation, and cline analysis was performed to explore the amount of gene flow between species and looking for drivers of introgression in bluebells. The evolutionary history of both species in northern Spain has been assessed using coalescence simulations.

In this study non-genetic evidence of hybridisation (chapter 2) will be complemented by analysing the genetic marker set developed in chapter 3. In particular, the following four questions were addressed:

1. What resolution do the SNP markers provide and are they biased by their design?
2. Where is the centre of the hybrid zone and are there barriers to gene flow?
3. Is there genetic evidence of heterosis-driven introgression in the hybrids?
4. Is the hybrid zone a consequence of secondary contact or primary intergradation?

4.2 Material and Methods

4.2.1 Sampling along the hybrid zone and molecular markers

For population genetic analyses of the hybrid zone, 311 samples from 48 collecting sites were included. This included samples from the field work as well as specimens from the museum accessions (Table A.1). Because of the strongly disrupted hybrid zone's geography, and the morphological characterisation of two hybrid races, they were treated separately as hybrid North, hybrid South and along with *Hyacinthoides non-scripta* and *H. hispanica* as the parental populations.

Isolating DNA and re-sequencing of the designed markers from genomic DNA, and subsequent variant discovery was performed following the methods outlined in chapter 3. From variant discovery bi-allelic single nucleotide polymorphisms (SNPs) were extracted and transformed into genotypic data, for which '0' denotes homozygous state for the *H. non-scripta* allele that is most frequent in the pure *H. non-scripta* individuals, '2' denotes homozygous state for the alternative allele, and '1' denotes heterozygous state at a SNP position.

The organelle (mitochondrion, and plastid) genomes were assumed haploid per individual and treated separately from the nuclear variant discovery. As shown in chapter 3, only two organelle haplotypes for the samples from the hybrid zone were discovered (concordance of mitochondrion and chloroplast). The same haplotypes were recovered in the

larger set of samples in the hybrid zone area presented here. Therefore, the organellar markers were not included in the population genetic analysis, but they are useful to determine the maternal lineage in hybrids since bluebells inherit their organelles maternally (Sears, 1980).

The vcftools v0.1.12b (Danecek et al., 2011) was used to obtain a ratio of transitions to transversions ($t_i t_v$), and nucleotide diversity (π). More specifically, for π a window size of 10 bp from the first observed variant to the last covered base was used to include non-variant sites and the estimates averaged over all windows. Most data manipulations and analyses were performed in R version 3.3.1 (R Core Team, 2016), unless stated otherwise.

4.2.2 What resolution do the SNP markers provide and are they biased by their design?

Resolution of genetic markers to differentiate parental and hybrid individuals (PCA and AMOVA). Initial analyses aimed at understanding the main source of variance for all biSNPs. Therefore, principal component analysis (PCA, *prcomp()*, package stats), and an analysis of molecular variance (AMOVA, *poppr.amova()*, Excoffier et al. (1992), package ade4 and poppr v 2.2.0 by Kamvar et al. (2014, 2015)) were performed. For the PCA missing data was imputed prior to the analysis (*imputePCA()*, Josse and Husson (2016), package missMDA).

The AMOVA was used to explore variance in allele frequencies at different hierarchical levels (i.e. individuals, collecting sites, and populations) and also to test the justification of considering four populations in contrast to three (i.e. merging collecting sites into populations that separated the hybrids into two different populations). Incompletely called variants were ignored in the AMOVA. The significance of the hierarchical clustering applied in AMOVA was addressed by permutation test (*randtest.amova(),* Excoffier et al., 1992). Thereby, the AMOVA is repeated with randomized assignment of an individual's origin for 1000 times. A null distribution for each variance component is obtained by assuming all samples from a global population. Individuals were then randomly drawn and allocated to a randomly chosen population. The sample sizes within respective level (i.e. collecting site, population) of the observed data are maintained. The variance components were estimated for each of the 1000 permutations and used as panmictic distribution of covariances to test the observed variance components for significant deviation. See R documentation of how p-values were estimated in *randtest()*: 'If the alternative hypothesis is greater: (number of random values equal to or greater than the observed one + 1)/(number of permutations + 1). The null hypothesis is rejected if the p-value is less than the significance level of $\alpha = 0.05$'.

Genetic differentiation by F-statistics. The amount of genetic differentiation (F_{ST}) due to genetic structure was calculated using Wright's F-statistics (Wright, 1978). To simplify explanations, I will henceforth only refer to the estimation of total F_{ST} between all four populations. Nevertheless, estimates of F_{ST} for subsets of data and for a certain level were estimated likewise. The observed heterozygosity, Ho, is the frequency of heterozygous individuals at a given locus in a certain group (collecting site, population, or full data set).

The expected heterozygosity, He, is the frequency of heterozygous individuals in panmixia (Hardy-Weinberg equilibrium). The fixation index (F_{ST}) is calculated as

$$F_{ST} = 1 - \frac{H_S}{H_T} = 1 - \frac{\sum_i \frac{n_i}{\sum_i n_i} 2 * \bar{p}_i * (1 - \bar{p}_i)}{2 * \bar{p} * (1 - \bar{p})}, \tag{4.1}$$

with H_S the expected heterozygosity within each population averaged over all populations (weighted by their size of n_i individuals) and H_T, the heterozygosity in the whole of four populations in absence of any structure (compare Nei and Chesser, 1983). $\bar{p}$ is the mean allele frequency over all, or per individual indicated by subscript ($_i$). Significance of F_{ST} was assessed by randomisation of the sample assignment into a population (1000 permutations). P-values were assigned after Bonferroni correction for multiple testing ($\alpha = 0.05$).

Inbreeding coefficient for a collecting site was estimated by $F_{IS} = 1 - Ho/He$, and 95 % confidence intervals were obtained by bootstrapping over individuals with 1000 simulations. Significance of F_{IS} was assessed by 1000 permutations of the individual's haplotype: For a given population or collecting site, at each locus the different possible alleles are written, i.e. either 0 or 1. Genotypes were randomly formed and used to generate the distributions of likely F_{IS} given the allele frequency. P-values were obtained after Bonferroni correction ($\alpha = 0.05$). Significant deviations from the expectations may indicate either population sub-structure, non-random mating or clonal reproduction ($F_{IS} > 0$); or excess of heterozygosity caused for instance by inbreeding ($F_{IS} < 0$). All these measures were averaged over all loci for desired levels (collecting site or population).

Isolation-by-distance. Isolation-by-distance (IBD) was tested between collecting sites within each population because it provides an alternative explanation to population differentiation by non-random mating. Especially since bluebells exhibit limited seed dispersal, pollen transfer is probably the main mode of gene flow. The pairwise haversine distances between sites (*distHaversine()* of R package geosphere; Shumaker and Sinnott, 1984) were estimated and compared to the pairwise F_{ST} between collecting sites by performing a Mantel test with 100,000 permutations (*mantel.rtest()* of R package ade4; Mantel, 1967; Thioulouse et al., 1997).

Introduced bias by marker design? The markers were selected using the criterion that the transcriptomes from *H. hispanica* (one individual BB-339) and *H. non-scripta* (one individual BB-411) were homozygous for different alleles. This rule was intended to enrich the dataset for alleles, which had substantial differences in allele frequency between the *H. hispanica* and *H. non-scripta* populations (see chapter 3). Consequently, they were likely to generate heterozygous markers in the hybrids.

This marker design potentially biased the F_{ST} distribution against neutral markers. Therefore, we needed to evaluate whether the difference between the neutral expectation and the actual F_{ST} distribution can be explained just by the initial bias introduced by the SNP design. The effect was explored by first estimating the distributions of allele frequencies for simulated neutral variants that evolved under HW equilibrium and panmixia.

These were compared with their subset of simulated variants restricted to the marker design. Based on equation (5.19) in Charlesworth and Charlesworth (2010), the equilibrium distribution of neutral variants ($\phi(p, \theta)$) is given in equation (4.2). There, theta, θ, is the mutation rate per locus scaled by 4Ne and it was obtained from coalescence simulations and subsequent parameter estimations using approximate Bayesian computations ($\theta = 0.00419$, see supplement 4.7.2). p is the frequency of a given allele, and Γ is the Gamma function, which is an extension of the factorial function from integers to real numbers.

$$\phi(p, \theta) \approx \frac{\Gamma(2\theta)}{\Gamma(\theta)^2} p^{\theta-1}(1-p)^{4\theta-1} \tag{4.2}$$

The probability of choosing a neutral marker that is homozygous and different in the two reference individuals from the transcriptome is given in equation (4.3). You choose to pick a locus '00' in *H. non-scripta* with a frequency of p^2, which is '11' in *H. hispanica* with a frequency of $(1-p)^2$ (– or the other way '11' is *H. non-scripta* and '00' in *H. hispanica*, which explains why the previous term is taken times 2) and weight the obtained number by the likelihood of observing this allele of frequency p in the equilibrium distribution of equation (4.2).

$$P(p) = 2p^2(1-p)^2\phi(p, \theta) \tag{4.3}$$

From the two distributions of allele frequency ($\Phi(p)$ and $P(p)$) and assuming a single panmictic population, F_{ST} values were obtained from 1000 replicates each (see above). Exactly 307 individuals were simulated from each distribution and their density histograms plotted (Figure 4.5). Finally, these were compared with the F_{ST} density histogram for the actual observed data for the subset of targeted markers performing Kolmogorov-Smirnov tests (*ks.test()*; Marsaglia et al., 2003; R Core Team, 2016).

4.2.3 Where is the centre of the hybrid zone and are there barriers to gene flow?

Hybrid index. To assess the amount of admixture in the hybrids Bayesian inference of ancestry was performed for K clusters (K = 2, 3, 4), applying fastSTRUCTURE to the bi-allelic SNP data (Raj et al., 2014). Ten independent runs of the program for each K were performed and the run with least negative marginal likelihood was selected for further analyses. The co-ancestry analysis was used to infer a hybrid index, which represents the amount of genetic admixture from either parent in a hybrid individual (e.g. Walsh et al., 2016).

Inferring hybrid zone centre by a 1D transect. Previous sampling (chapter 2) showed that the transition from *H. non-scripta* to *H. hispanica* occurs mostly from north to south, but when following allele frequency changes along latitude *H. non-scripta* individuals appeared as far south as *H. hispanica* to the west of the hybrid zone (Figure 4.1). In addition, hybrid populations occur East-West on either side of the Sierra del Teleno

(Figure 4.1, see chapter 2 for detailed description). To linearise the spatial distance between collecting sites, the geographic coordinates were transformed into a 1D transect (Stankowski et al., 2016). Based on the samples' coordinates and their hybrid index, a surface distribution of admixture was extrapolated using the *Contourpointplot3D* function of Mathematica (v10.3.; Wolfram Research, 2016). The 0.5 isocline (i.e. complete admixture) was used as the centre ('0') of the hybrid zone. Distances from collecting site to the predicted hybrid zone centre were estimated as the shortest straight line in km (*dist2Line()*, package geosphere). Accordingly, a 1D transect was obtained that sorts the collecting sites by their distances to the 0.5 isocline with sites south of the centre (*H. hispanica*) showing negative distances and sites north of it (*H. non-scripta*) showing positive distances.

Diagnostic markers. For the cline analysis, particular interest was in loci, which showed substantial differences in allele frequency between *H. non-scripta* and *H. hispanica* taxa, i.e. diagnostic loci. Suitable loci were assessed by another run of fastSTRUCTURE. Pure parental samples with a mean admixture (Q) per collecting site of either < 0.01 or > 0.99 were selected and analysed for K = 2. Loci were chosen if the posterior mean allele frequency was above 0.6 in one species and below 0.4 for the other.

Fitting sigmoid curve to each locus and extracting cline parameters. Using generalised linear models (GLM), individual allele frequencies at each locus were regressed on the distance of each collecting site from the centre of the genome-wide 0.5 isocline (i.e the 1D transect). In depth, a logistic regression model with a quasi-binomial error distribution and a logit link function (*glm()*, package stats) was used to predict the allele frequencies across the 1D transect. The coefficients (a = slope, b = intercept) and their standard errors were extracted from each locus model. Using the function of the sigmoid curve

$$y = 1 - \frac{1}{1 + \exp(a * x + b)} \tag{4.4}$$

the centre and width of a cline were estimated. In equation (4.4) x stands for the position of a collecting site across the 1D transect and y for the predicted allele frequency at each locus given the obtained coefficients (a, b) from the model. Further, the estimates of P_{min}, P_{max}, and δ_P from the predicted cline curve were estimated. P_{min} and P_{max} stand for the allele frequency at either end of the cline, with δ_P as their difference. The cline centre (estimated as $-intercept/slope$) is the inflection point of the sigmoid curve and coincides with the steepest position of allele frequency change where the projected allele frequency is 0.5. The hybrid zone centre is an important measure of the locality of disrupted gene flow. The width is the inverse of the slope and reflects the strength of selection pressure on the allele frequency. Strong selection against gene flow, divergent selection, or a population density drought relative to the rate of gene flow would result in a steep cline and increase the reproductive barrier. All variants were selected that showed a slope coefficient significantly different from zero ($Pr(> |t|) \leq 0.05$) and whose centre estimate were within the 1D transect.

Alternative clines. The nuclear clines were compared to alternative evidence clines in regards to centre positions (coincidence), and steepness of their clines (concordance). Therefore, the hybrid index, morphological first principal component (which was rescaled to 0-1), and the organelle haplotypes (but with binomial link-function) were similarly fitted across the 1D transect.

4.2.4 Is there genetic evidence of heterosis-driven introgression?

In seed crosses evidence for high hybrid fecundity and frequent formation of F1 hybrids was found, which was interpreted as hybrid advantage assuming low post-zygotic selection on the hybrids based on their abundance in the field and reported elsewhere (chapter 2). The intermediate admixture proportion observed for hybrid North could indicate genome-wide genetic heterozygote advantage. Excess of heterozygosity in hybrids across the 1D transect was assessed by F_{IS} estimates per collecting site (section 4.2.2). To pursue the hypothesis of advantageous heterozygous alleles in hybrids that maintain clines but also drive introgression, the cline fitting was repeated as described above but removing in turn each of the hybrid population. The slope, and cline centre along with their standard errors (representing the fit of a cline to individual allele frequencies) for the hybrid population remaining in the analysis were compared.

The hypothesis was made that heterosis-driven introgression in more admixed hybrid North due to higher frequency of heterozygous genotypes (GT = 1, results 4.3.3, Figure 4.6) would lead to less genetic differentiation (Lindtke et al., 2012) reflected by shallower slopes and their standard errors would be larger because the residuals to the fitted clines are larger. In contrast, the clines of less heterozygous hybrid South would be steeper (GT = 0, or GT = 2) and show smaller standard errors.

However, just by removing one of the hybrid populations a shift of the centre position and steepness of slope would be introduced, which biases the results. Therefore, the effect of removing a population from the data was explored by simulating clines, which follow the allele frequencies in each collecting site of the observed data: We assume 48 sample positions that are distributed across the 1D transect (Table 4.8). Using the slope and cline centre of the hybrid index (section 4.3.3, Table 4.3) and equation (4.4), the allele frequency of each collecting site along the transect was predicted. Next, to generate individuals in this population by random draw, the frequency of the alternative allele in the collecting site was used as one of the two parameters. The other parameter corresponded to the number of possible haplotypes of the sequenced individuals in each collecting site. Random mating was assumed and that the two copies of the gene are independent.

Using the same GLM as above (i.e. logistic regression model with a quasi-binomial error distribution and logit link function), the slope and centre of the cline were estimated for the full data set, as well as the two subsets, one corresponding to all individuals except the hybrid North (referred to as HyS), and the second case corresponding to all individuals except the hybrid South (referred to as HyN). This re-sampling method was repeated 10^5 times and the distribution of the cline estimates were compared between the different subsets of included populations. Only simulations were kept that showed significant slopes for all three data subsets. The significance of a shift in slope and centre estimates was assessed by pairwise Kolmogorov-Smirnov tests (ks-test). And lastly, the

difference between the distributions with 95 % confidence intervals was estimated, and it was tested if the difference is significantly different from zero using one-sample t-test (*t.test()*, package stats).

4.2.5 Is the hybrid zone a consequence of secondary contact or primary intergradation?

The evolutionary history between the parental populations was mostly addressed using coalescent simulations and approximate Bayesian computation for parameter estimation of effective population size, species split between both parent populations, and migration from the most likely population history model. Since the analyses must be regarded preliminary at this point, they were only included as section in the supplementary (section 4.7.2). They are still mentioned here because part of the results (the mutation rate θ) were used in the inferences of the marker bias section (section 4.2.2).

4.3 Results

4.3.1 Data sampling and molecular markers

Amount of obtained SNPs. Four samples were excluded due to failed amplicon resequencing and admixed parental collecting sites at the margins of the hybrid zone, leaving for analysis 307 individual bluebells collected in Northern Spain. They were obtained from 48 different collecting sites in the provinces Castille and León, and Galicia (Figure 4.1). This data set included 81 samples of *Hyacinthoides hispanica* from 12 sites, 105 samples of *H. non-scripta* from 18 sites, 61 samples of hybrid North from nine sites, and 60 samples of hybrid South from nine sites.

After removing primer regions, the amplicons totalled 37,393 nucleotides with a mean amplicon length of 131.5 bp. There were 4343 variants, of which 42.8 % passed the stringent quality filters. If a variant site failed calling in more than a third of the samples due to low read coverage, the variant was removed from analysis. The passed variants contained a marginal 2.9 % of structural variants (e.g. insertions or deletions). Single nucleotide polymorphisms (SNPs) accounted for 39.9 % variants, of which 0.9 % showed more than two alleles.

All 1,640 bi-allelic SNPs (biSNPs) were found in 215 different genes with 1 to 17 biSNPs per amplicon. The majority of these SNPs (54.57 %) were rare alleles with a minor allele frequency below 5 %, including 18.23 % as singletons. The mean overall nucleotide diversity was 0.0193 (Table 4.1) and the transition to transversion ratio of biSNPs was 2.604. When comparing the biSNPs to the two parental transcriptomes, the majority of biSNPs, 72.99 %, occurred in new positions, and 15.91 % were target markers and polymorphic across 307 samples (261 of 361 designed targets – 72.3 %). Consequently, the remaining 11.1 % of biSNP positions were also polymorphic between the two transcriptomes. Focusing on the subset of parental samples, only 246 target sites remained polymorphic, of which in turn only 6.5 % (16 of 246) of target variants were near fixation (i.e. allele frequency either less than 0.2 or more than 0.8 in either population, respectively). However, there were 24.39 % of target SNPs that represented private alleles

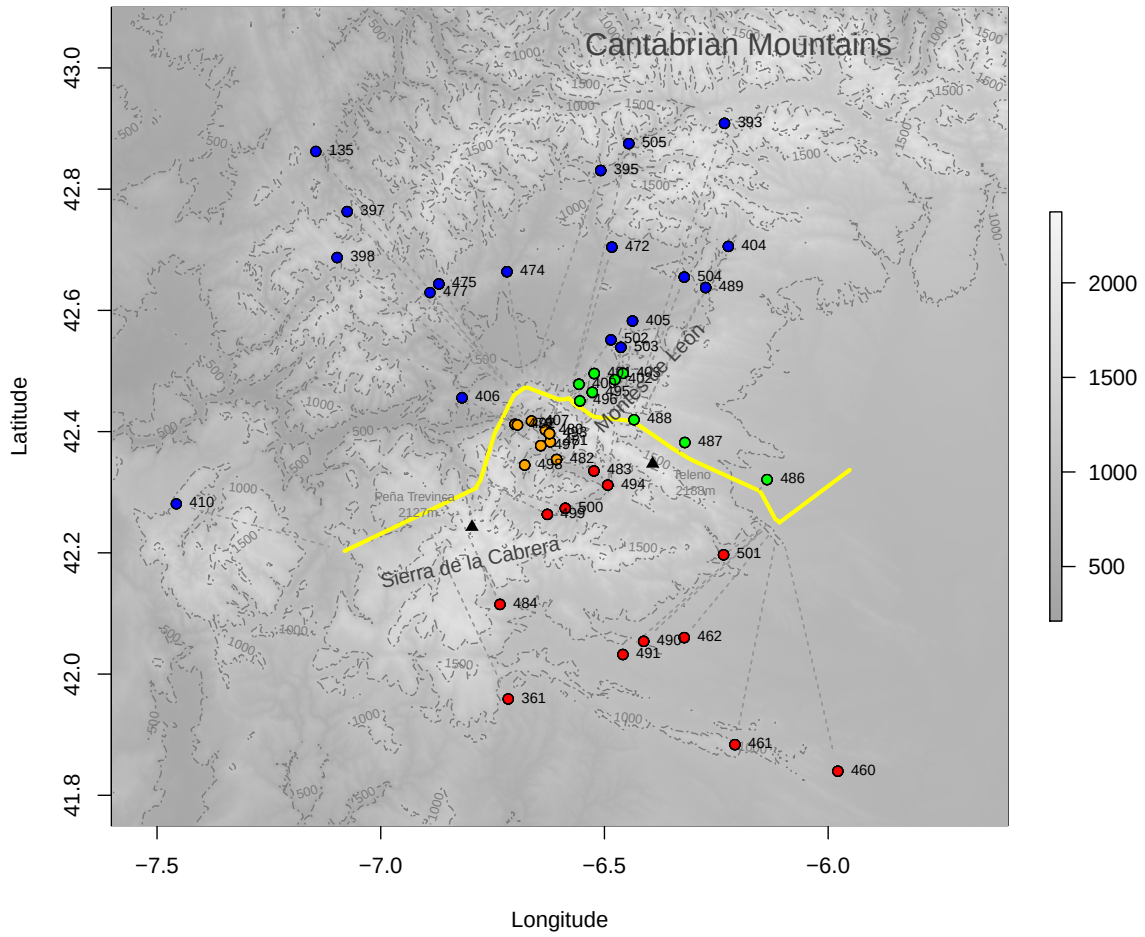


Figure 4.1 – Topographic contour map of study area including the inferred 1D transect as cline centre (yellow) and the shortest distances to collecting sites (grey dashed lines). Each dot represents a collecting site with its identifier. The colour indicates the identified taxon from the field: blue – *H. non-scripta*, green – hybrid North, orange – hybrid South, and red – *H. hispanica*.

(i.e. alleles restricted to a parental population). All other target positions (69.11 %) were shared polymorphisms at intermediate frequencies.

4.3.2 What resolution do the SNP markers provide and are they biased by their design?

Inter-population variation. All 1,640 biSNPs showed a strong differentiation between the four populations in the principal component analysis (PCA, Figure 4.2). The first principal component (PC1) widely spaced either parental species and explained 20.4 % of the observed variance in the data. The second component (PC2) accounted for 1.8 % of the variance, which spread the hybrid South and *H. hispanica* individuals. The remaining principal components only explained less than 1.61 % each, and they did not cluster the samples into any recognisable pattern. Further, there was a clear separation between *H. non-scripta* and hybrid North (apart from a few *H. non-scripta* samples from BB-406), in contrast to *H. hispanica* and hybrid South, for which the 95 % confidence ellipses overlapped greatly (Figure 4.2). Some supposedly southern hybrids (from BB-482) fell within the range of supposedly pure *H. hispanica* (BB-483, BB-494, BB-499, BB-500). *H. hispanica* individuals that were most distant to the hybrid zone centre (BB-460, BB-461, BB-490, BB-491, Table 4.8) clustered outside of the overlap with hybrid South (at PC2 around -4). Lastly, one outlier (BB-479-10) was observed that overlapped with *H. hispanica* in several principal components. Excluding this sample from PCA did not largely affect the overlap of the 95 % ellipses between hybrid South and *H. hispanica* (not shown). The overlap of hybrid South and *H. hispanica* raised concerns about the purity of collected *H. hispanica* samples.

Using an analysis of molecular variance (AMOVA, technically covariances) the justification of considering four populations was tested; i.e. that there were two ancestral populations and that the hybrids should be split in two populations, one northern and one southern. Two AMOVAs were performed that tested the variance of the samples grouped into collecting sites and these into either three ('Pop3') or four ('Pop4') populations (supplement table 4.7). In both cases, most of the covariance was observed within individuals (73.09 % and 74.02 %) followed by variation between populations (15.86 %, and 16.06 %, respectively). The variation between collecting sites within populations were smaller for four populations than for three, which was expected when fewer collecting sites were compared per population. The covariance between populations was non-significant for both analyses (supplement table 4.7). Thereafter, the collecting site level was removed and the analyses repeated. For four populations, (as well as for three populations – not shown), a significant covariance of 16.4 % between populations was obtained (see 'Total' in Table 4.1).

It was also tested, if the variation between the two hybrid populations was quantifiable by including only the hybrid samples and population level. The contribution of variation between hybrid populations was significant (σ = 8.3, 6.9 %, $p \leq 0.001$). For comparison, the estimated contribution to variations between the parental populations was much larger (σ = 42.9, 29.7 %), and the 'back cross' between *H. non-scripta* and hybrid North (σ = 8.4, 7.2 %), as well as between *H. hispanica* and hybrid South (σ = 8.4, 8.0 %) were in the

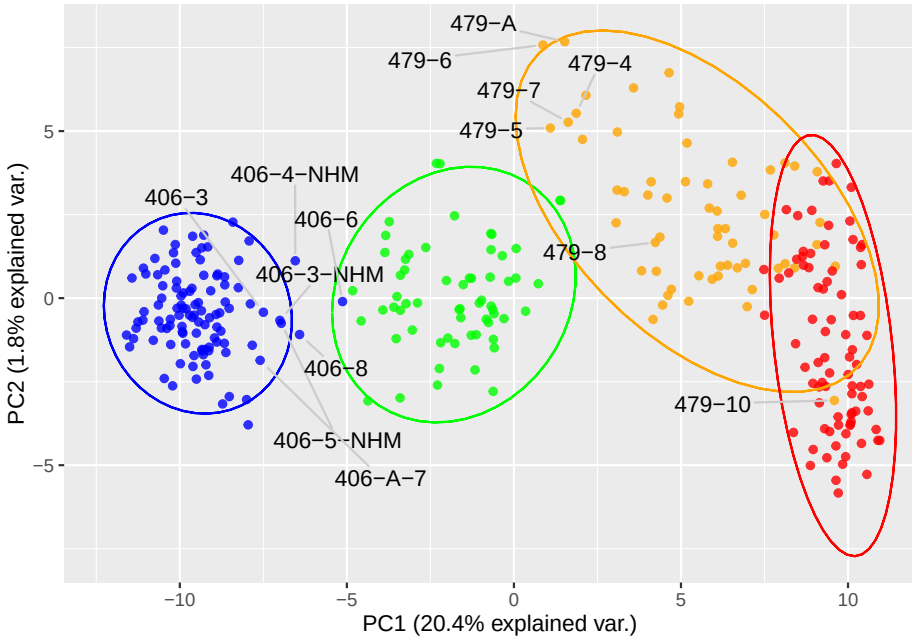


Figure 4.2 – First and second principal components of 1,640 biSNPs for each individual. For each population, the 95 % confidence ellipses were plotted obtained under the assumption of bivariate normal distribution. The populations are coloured with: blue – *H. non-scripta*, green – hybrid North, orange – hybrid South, and red – *H. hispanica*.

realm of between the two hybrid populations. The AMOVA results were congruent with the PCA in that more variation between individuals was observed within a population; rather than between their collecting sites within a population. The AMOVA also provided evidence to treat the hybrids as two separate populations; as did the PCA, in which the two hybrid populations were clearly separated by PC1.

The overall genetic differentiation (F_{ST}) of the total population was small (0.062, Table 4.1), but the pairwise difference (Table 4.2) between both parental species was larger ($F_{ST} = 0.083$). The hybrids showed lowest pairwise F_{ST}. The difference between hybrid North and *H. non-scripta* ($F_{ST} = 0.025$) was larger than between *H. hispanica* and hybrid South ($F_{ST} = 0.018$), which is congruent with the PCA result. Therefore, pairwise F_{ST} is increasing the farther apart the populations are. Interestingly, the difference between both hybrid populations was almost as large as between *H. non-scripta* and hybrid North ($F_{ST} = 0.024$). When selecting the target SNPs that were polymorphic in the respective subset of samples, the pairwise F_{ST} estimates were about doubled between the parents but not between the hybrids (Table 4.2). This indicated that the marker design was successful in selecting diagnostic alleles that detect genetic differentiation between parents and that segregate in the hybrid individuals.

Intra-population variation between collecting sites. The population F_{ST} estimates (between collecting sites within population) were smaller for the hybrid populations than for their parents, but they were also collected over a smaller range and included fewer samples/sites (Table 4.1). Within each population 87 – 91 % of the covariance was due to variation within individuals (Table 4.1). For hybrid South the covariance between samples within collecting sites was greater than the covariance between sites. The opposite was

Table 4.1 – Summary table for four populations with the results of genetic differentiation using F- and ϕ-statistics (obtained from AMOVA). For the four populations their collecting sites were taken as sub-level, while for the 'total' estimate individuals were compared to population level directly. The component of covariance (σ) and their contribution to the total covariance (%) are also reported. Significance of covariance was assessed using 1000 permutations with: $p \leq 0.05$ (*), $p \leq 0.01$ (**), and $p \leq 0.001$ (***).

	H. non-scripta	hybrid North	hybrid South	*H. hispanica*	total
Samples	105	61	60	81	307
Sites	18	9	9	12	48
π	0.022	0.025	0.023	0.021	0.019
Ho	0.113	0.132	0.119	0.108	0.117
He	0.130	0.147	0.136	0.125	0.154
F_{IS}	0.132	0.108	0.112	0.127	0.184
F_{ST}	0.137	0.102	0.104	0.124	0.062
σ bt sites	6.671^{***}	5.297^{***}	5.104^{***}	7.687^{***}	20.921^{***}
σ bt samples wi sites	4.794^{*}	$4.870^{0.062}$	6.991^{***}	4.685^{*}	12.144^{***}
σ wi samples	91.306^{*}	106.464^{***}	95.321^{***}	86.888^{***}	94.174^{***}
σ total	102.771	116.630	107.415	99.260	127.240
% bt sites	6.491	4.542	4.751	7.744	16.442
% bt samples wi sites	4.665	4.176	6.508	4.720	9.544
% wi samples	88.844	91.283	88.741	87.535	74.014

Table 4.2 – Mean pairwise F_{ST} per locus contrasting the populations. Two subsets of markers are presented: Upper half target subset, lower half all loci.

	ns	hyN	hyS	hisp
ns		0.041	0.109	0.169
hyN	0.025		0.037	0.083
hyS	0.059	0.024		0.031
hisp	0.083	0.045	0.018	

the case for the remaining three populations. For hybrid North the variation between individuals within sites was actually not significant. Therefore, more genetic variance was observed within collecting sites in hybrid South than for hybrid North, for which it was the lowest. However, all populations showed significant variations between collecting sites; which was also lowest for both hybrid populations compared to the parental populations.

The inbreeding coefficient, F_{IS} for each population was positive ($F_{IS} > 0.1$) suggesting population subdivisions, but lowest for hybrid North (Table 4.1). Per collecting site the F_{IS} ranged from -0.15 to +0.29 with the majority of collecting sites showing slightly positive F_{IS} and significant deviation from zero (72.9 % positive with $p < 0.001$ after Bonferroni correction, their mean = 0.008, and median = 0.0056, Table 4.8). This indicates that averaged over all loci, collecting sites are close to Hardy-Weinberg equilibrium with a tendency towards heterozygotes deficits. All collecting sites with insignificantly deviating F_{IS} were negative. The strongest deviation from zero F_{IS} occurred in collecting sites with less than six samples. For collecting sites with only three samples the 95 % confidence intervals (obtained by bootstrapping) showed very poor estimates of F_{IS} (BB-135, BB-410, and BB-477, Figure 4.12). The highest positive F_{IS} was observed for BB-135, of which two samples have amongst the highest proportion of no-calls (10.9 % and 7.4 %)

and also lowest number of heterozygous sites. The strong positive F_{IS} for this collecting site might be due to poor DNA material in two out of the three samples and consequently failure of re-sequencing both alleles per amplicon locus in sufficient coverage for variant calling (see section 4.3.2).

Genetic isolation within each population between collecting sites (pairwise F_{ST}) due to geographical distance was tested (Figure 4.3). On the one hand, the parent populations showed a significant correlation, which was stronger for *H. hispanica* (r = 0.65) than for *H. non-scripta* (r = 0.52). However, the maximum pairwise F_{ST} was highest for *H. non-scripta*. The hybrid populations, on the other hand, showed weaker isolation by distance (IBD). Across a distance of 50 km hybrid North showed no correlation at all, yet hybrid South presented a small (r = 0.39) but still significant correlation across 12 km. This relates to the AMOVA result that hybrid North showed the lowest variance component between individuals within collecting site and between collecting sites (Table 4.1). However, hybrid North presented the highest genetic diversity (π) and heterozygosity (Table 4.1). Isolation by distance was also tested, if the hybrids were treated as one population. Together they showed similar correlation as hybrid South but not strongly significant (N = 630 (18); r = 0.314, P = 0.044). Overall, the pairwise geographic distance estimates might underestimate the real spatial distance, which was measured 'as the crow flies' and therefore ignores intervening terrain like the mountain peaks between both hybrid populations.

The observed IBD for all but northern hybrids fits with the observation of AMOVA and by population F_{ST}, which showed stronger differentiation between collecting sites and within individuals leading to positive F_{IS} per population (Table 4.1). This strong level of population subdivision can also explain the reduced total F_{ST} estimate as the total variance is partitioned to within and between individuals. The lack of heterozygosity within collecting sites could be explained by higher relatedness due to 1) limited gene flow between sites, 2) clonal reproduction, and 3) non-random mating such as caused by cyto-nuclear incompatibilities.

Introduced bias by marker design? The marker design aimed at species specific variants between *H. non-scripta* and *H. hispanica* from exon sequences, and in Table 4.2 it showed that the target SNPs provide stronger genetic differentiation. In a panmictic population, most of these targeted markers would be expected to have a frequency near 0.5, which is the rarest form of neutral alleles when the mutation rate (θ) is small (Figure 4.4).

To evaluate the impact of the SNP selection process on overall genetic differentiation in neutral markers three different F_{ST} distributions where compared: density distribution of F_{ST} for neutral simulated markers in general, simulated neutral markers selected under the marker design, and the selected markers in the observed data (Figure 4.5).

The result was a slight positive shift of mean F_{ST} density between expected neutral markers and the F_{ST} distribution of neutral markers as designed (Figure 4.5) with significant differently shaped distributions (two-sample Kolmogorov-Smirnov test: D = 0.317, p-value < 0.001). Hence, the marker design did modify the distribution of neutral markers. But the same holds for their comparison to the actual observed data. Indeed, the observed F_{ST} of target SNPs is 30-fold larger (also see Table 4.2). To conclude, the introduced bias

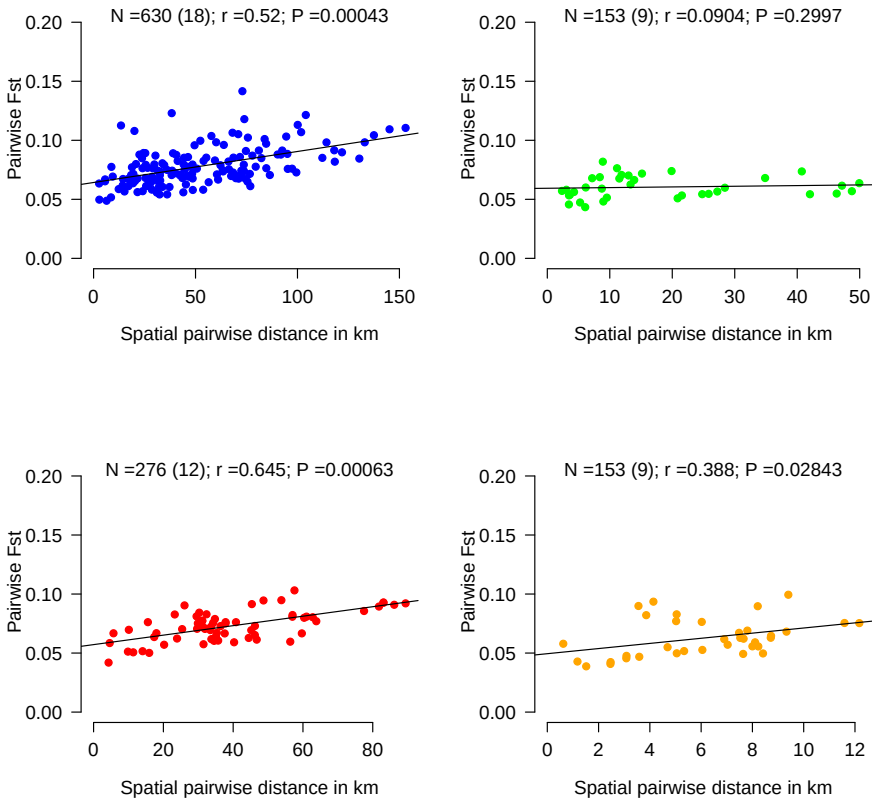


Figure 4.3 – Linear regression coefficient plotted of pairwise genetic differentiation (F_{ST}) against spatial distance between collecting sites. Estimates were obtained by N pairwise comparisons of (x) collecting sites for each population (blue – *H. non-scripta*, red – *H. hispanica*, green – hybrid North, orange – hybrid South). Mantel-test (correlation coefficient r; permutation test) was performed to predict p-values.

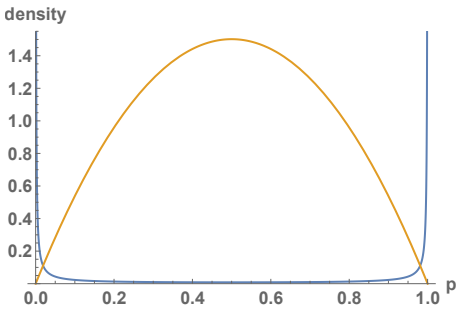


Figure 4.4 – Expected (density) distribution of allele frequencies in a panmictic population with mostly rare alleles – blue, and of allele frequencies of the selected markers that are homozygous in the reference plant (00, 11), but are polymorphic loci (00, 01, 11) with intermediate allele frequencies in a panmictic population – orange. They correspond to equations (4.2) and (4.3), respectively. The mutation rate, $\theta = 0.00419$, obtained from ABC was used here.

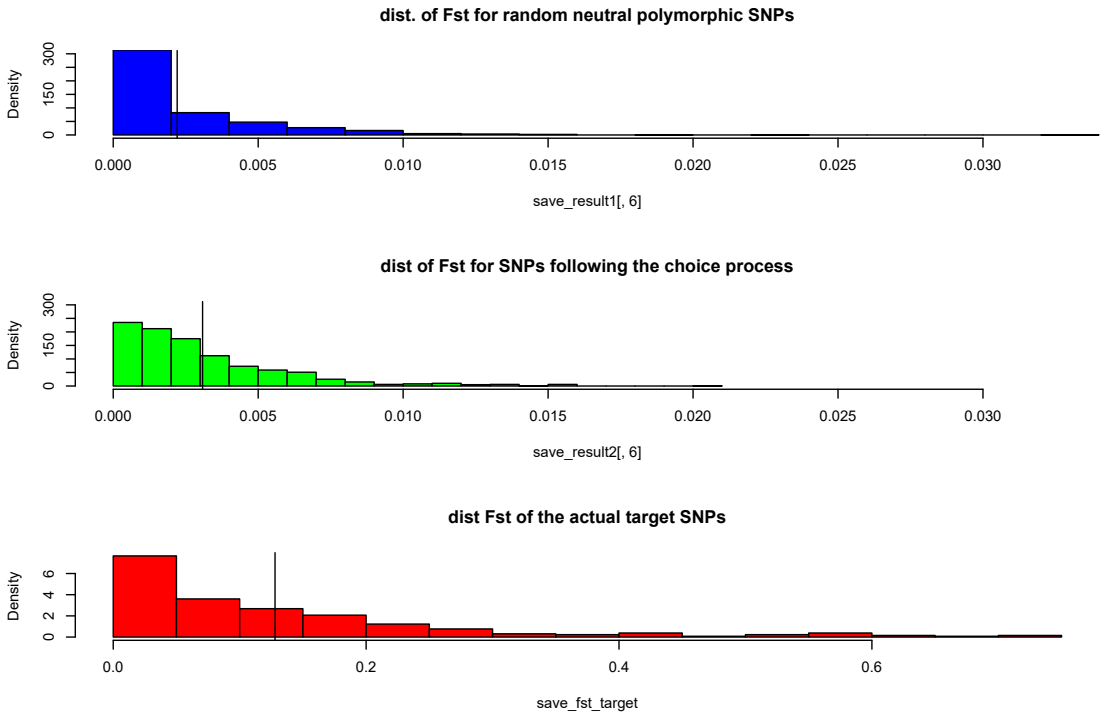


Figure 4.5 – Density distributions of F_{ST} for any neutral marker – blue, neutral markers using the marker design – green , and the actual observed F_{ST} in the subset of the target SNPs – red. The black line is the mean of each F_{ST} density distribution. The distributions are all significantly different from each other in shape as tested by two-sample Kolmogorov-Smirnov test: blue vs. green (D = 0.337, p-value < 2.2e-16), blue vs. red (D = 0.891, p-value < 2.2e-16), green vs. red (D = 0.879, p-value < 2.2e-16).

is insufficient to explain the F_{ST} distribution observed in the data, and F_{ST} therefore reflects real population structure, and is not simply generated by the marker design.

4.3.3 Where is the centre of the hybrid zone and are there barriers to gene flow?

Hybrid index. For all fastSTRUCTURE runs, K = 2 provided the highest marginal likelihood and explained most of the structure across all bi-allelic SNPs (Figure 4.6). Each cluster could be assigned to either parental genome's of *H. non-scripta* (meanQ = 0.99 ± 0.034), and *H. hispanica* (meanQ = 0.0072 ± 0.019). The admixture proportion of *H. non-scripta* per individual was therefore used as the hybrid index (100 % is *H. non-scripta*), which quantifies the contribution of either parental genome in hybrid individuals. The northern hybrids showed more intermediate admixture proportions (meanQ = 0.61 ± 0.076) than the southern hybrids (meanQ = 0.22 ± 0.12) (Figure 4.6). Some supposedly pure parental collecting sites (field observation) showed admixture with the other parent. Especially among *H. hispanica* the mean per collecting site ranged between 1 - 4 % of admixture for four sites closest to the range of hybrid South (BB-483, BB-494, BB-499, BB-501; Table 4.8; Figure 4.6). This suggests that gene flow reaches further into *H. hispanica* than noted by observations during field collection. The BB-479-10 individual that clustered with *H. hispanica* in the PCA, occurred to be genetically 100 % of *H. hispanica* ancestry and is the exception amongst hybrid samples. For *H. non-scripta* one collecting site (BB-406), which is located at the west end of the hybrids' southern valley, presented considerable admixture of 13 % from *H. hispanica* (Figure 4.2). Collecting sites BB-502 and BB-503 of *H. non-scripta* showed small admixture proportions (Table 4.8), which might be explained by their proximity to northern hybrids. Site BB-135 in the distant north-west of the study area presented 5 % admixture, but based on its distant position to the hybrid zone centre, gene flow from hybrids or *H. hispanica* seems implausible and it could possibly be due to technical errors as discussed in section 4.3.2.

Diagnostic markers. For the cline analysis of genetic loci, the interest was particularly on biSNPs with diagnostic allele frequencies in either species (i.e. at least 0.6 difference between parental taxa). Re-running fastSTRUCTURE for 142 samples of pure origin (i.e. 60 *H. hispanica* samples and 82 *H. non-scripta* samples) and assuming two ancestral clusters, the posterior allele frequency was extracted for each biSNP in the respective cluster. Each cluster could be assigned to either *H. non-scripta* or *H. hispanica* (Figure 4.7). The probability of allele frequencies per biSNP showed mostly low or high frequencies for both clusters, which indicates a large proportion of rare alleles as well as shared polymorphisms (Figure 4.7). Only 70 SNPs were found in 43 different genes (45 different amplicon sequences) that matched the diagnostic criterion. Fixed or highly differentiated markers between both genomes are therefore rare in this marker set, as already noted in section 4.3.1.

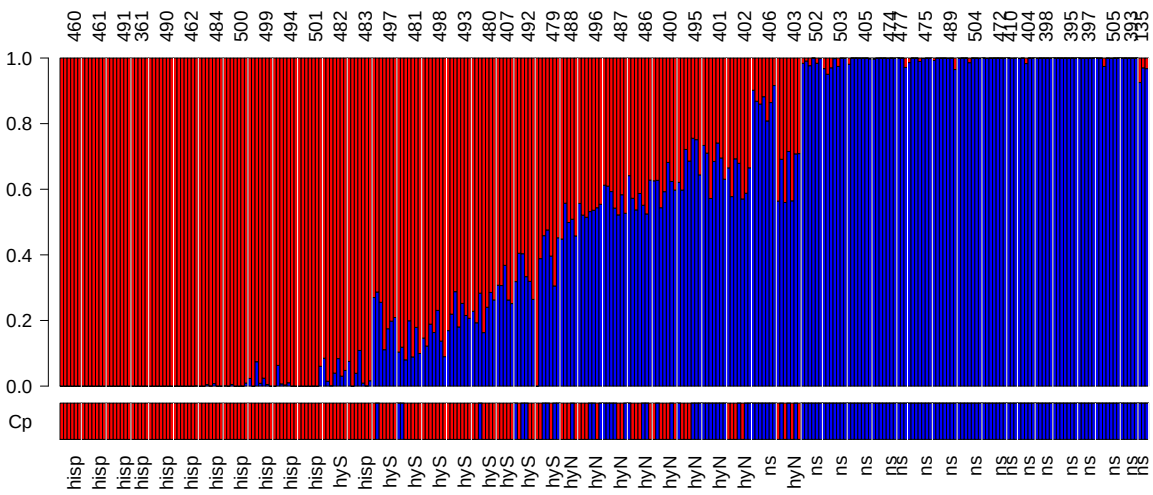


Figure 4.6 – Co-ancestry plot for K = 2. Samples ordered by distance to cline centre (between 479 and 488) with red representing *H. hispanica* genome and blue *H. non-scripta* genome proportion. Bottom barplot (Cp) shows the organelle haplotype per individual.

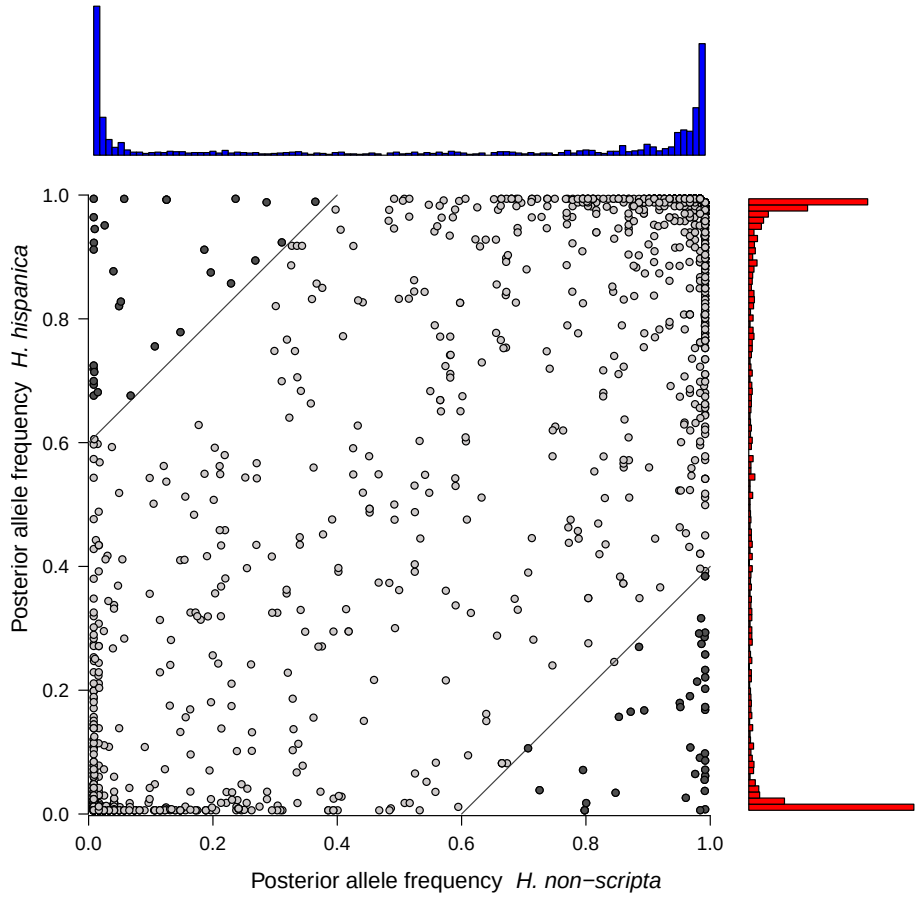


Figure 4.7 – Posterior mean frequency of the alternative allele of SNPs in each of the two inferred clusters for 142 pure parental individuals obtained from fastSTRUCTURE. The two histograms correspond to the distribution of allele frequencies in each parental population: blue – *H. non-scripta*, red – *H. hispanica*. Highlighted in black are the 70 loci that have diagnostic frequencies, which were used to differentiate the parental populations.

The alternative clines in relation to the hybrid zone centre. The centre of the hybrid zone – inferred from the 0.5 isocline of a surface distribution of nuclear admixture – occurred mostly north to the mountain ridge of the Sierra del Teleno (Figure 4.1). The cline inferred from all 307 samples' admixture proportions (i.e. hybrid index, 1 = *non-scripta* like) can be seen as the expected steep transition for loci that differentiate both parental genomes (Figure 4.8: left). Its centre was shifted towards *H. non-scripta* by 1.14 km but its 95 % confidence interval ranged from -1.4 km to +7 km (Table 4.3). The cline represented a significant fit to the samples with low standard error and it also shows that the 0.5 isocline provided a good approach to linearise the collecting sites in a 1D transect. A shift in nuclear clines from the hybrid index centre towards *H. non-scripta* can be interpreted as introgression of *H. hispanica* alleles towards *H. non-scripta*, and vice versa.

The organelle cline was steeper and shifted further towards *H. non-scripta* by +2.29 km. The 95 % confidence interval of its cline was narrower than the hybrid index cline, but it falls within its confidence interval. The overall mean frequency of *H. non-scripta* organellar haplotype in hybrids was low (0.35) showing that the majority of hybrid individuals bear the *H. hispanica* organelle haplotype. For hybrid North at least half of individuals bear the *H. non-scripta* haplotype (mean = 0.51), with three of nine collecting sites where the average was larger than 0.5 (BB-401, BB-487, BB-495, Table 4.8). The most northern *H. hispanica* organelle haplotype was found very close to *H. non-scripta* sites (BB-403, Figure 4.6). For hybrid South fewer individuals had the *H. non-scripta* haplotype (mean = 0.18) with two out of nine collecting sites where the frequency of *H. non-scripta* haplotype was near 0.5 (BB-479, BB-492, Table 4.8). These two collecting sites are closest to the next *H. non-scripta* locality, BB-406, which also showed considerable nuclear admixture but no *H. hispanica* organelle haplotype. The organelle haplotype moves very slowly with seed dispersal, and consequently, they can remain in their locality while the rest of the genome is replaced by introgression of nuclear markers – especially under pollen-driven migration.

The morphological cline provided a strongly significant fit with the lowest standard error, was shallower than the previous two clines and shifted towards *H. hispanica* by -2.43 km (Table 4.3). The shift was significantly different from the other two cline centres because its centre occurred outside of their confidence intervals. The morphology centre coincided with the transition between hybrid South and hybrid North (between BB-479 and BB-488).

Table 4.3 – Parameter estimates for the clines of hybrid index (HI), organelle haplotype (Org), and morphology (Mor). The estimates are obtained from the GLM with units for slope (km^{-1}), centre (km distance from zero), and width (km).

	Slope	**95% CI**	**StdErr**	**P-value**	**Centre**	**95% CI**	**Width**
HI	0.2068	[0.13, 0.314]	0.0464	$< 1.16e-05$	1.1383	[-1.358, 7.013]	4.8346
Org	0.2383	[0.182, 0.308]	0.0319	$< 8.32e-14$	2.2908	[0.451, 5.354]	4.1956
Mor	0.1186	[0.099, 0.14]	0.0104	$< 2.9e-19$	-2.4318	[-3.628, -0.741]	8.4325

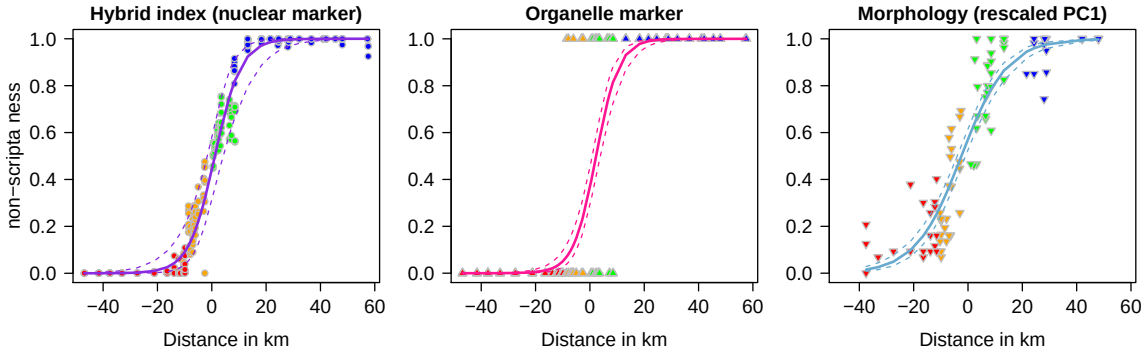


Figure 4.8 – Clines (solid lines) of hybrid index (left), organelle haplotype (middle), and morphology (right) and their 95 % confidence intervals (dashed lines). For details on the parameters see Table 4.3. The raw data points for each individual are plotted with their field identification indicated by red – *H. hispanica*, orange – hybrid South, green – hybrid North, blue – *H. non-scripta*.

The nuclear clines are shifted towards *H. non-scripta*. There was a large variation in cline positions and slopes between nuclear biSNPs. Regarding the nuclear markers, 1,002 out of 1,640 biSNPs have shown a slope significantly different from zero ($p \leq 0.05$) but only 316 of these also had their centre position within the region of hybrid individuals (-10.31 km to 8.53 km, Figure 4.9). These 316 SNPs occurred in 150 different nuclear genes (and 178 different amplicon sequences). The majority of biSNPs had shallow clines (median slope = 0.043 km^{-1}, mean slope = 0.049 km^{-1}). The 70 diagnostic alleles ('pure' parental samples of AF > 0.6 in one parent, and AF < 0.4 for the other parent; see section 4.2.3) largely overlapped with SNPs showing a $\delta_P \geq 0.832$ and their centre positions occurred within the region of hybrid individuals across the 1D transect (Figure 4.9). The steepest clines were observed in the range of northern hybrids and not in the middle of the hybrid cline. Hence, in contrast to the inferred 0.5 isocline from section 4.2.3, the nuclear markers were shifted further towards *H. non-scripta* (mean centre = +6.13 km, median centre = +8.5 km) than the hybrid index cline. There was no peak in centre positions along the transect (centre histogram of Figure 4.9), though. As the steepest clines are due to strong selection, they are most likely to escape the hybrid zone quickly; unless pre- or post-zygotic incompatibilities play a role and gene flow is limited.

Candidate genes for restricted introgression. Selecting the top 25 % of obtained slopes (i.e. steeper than 0.057 km^{-1}), 79 biSNPs in 43 different loci were found. They all showed significant F_{ST} (range 0.09 – 0.724, mean 0.386), and 85 % also showed significant and positive F_{IS} (range 0.31 – 0.84, mean 0.55) between the four populations. Consequently, these SNPs can contribute to genetic differentiation (strong F_{ST}) and potential lack of heterozygotes in the centre (postive F_{IS}).

Three SNPs (in two loci) had a steeper cline than the lower confidence interval of the organelle slope estimate. In addition to the concordant slope, they also coincided with the organelle cline centre defined as the 95 % CI of the centre estimate (Table 4.3, Figure 4.10). Another 11 biSNPs (in 10 loci) coincided with the organelle centre but with various slopes (range 0.065 – 0.29, mean = 0.15) and large δ_P (range 0.93 – 1, mean = 0.99).

In contrast, only three biSNPs (in two loci) fell within the range of the morphology cline centre and another three biSNPs (in two loci) were also as steep as its slope (Table 4.3, Figure 4.10). The biSNPs concordant with the morphology cline were similar in shape (slope: range 0.083 – 0.12, mean = 0.098; δ_P: range 0.97 – 1, mean = 0.98; centre: range -3.2 – -0.81 km, mean = -1.9 km).

Two other areas along the 1D transect presented a cluster of steep clines in the region of *H. non-scripta* (Figure 4.10). The first coincided with the transition from hybrid North to *H. non-scripta* between 5.354 to 13.22 km. There were 17 biSNP (in 12 loci) that presented a range of slopes (range 0.07 – 0.13 km^{-1}, mean = 0.089 km^{-1}), δ_P (range 0.94 – 1, mean = 0.97), and centres (range 6.9 – 12 km, mean = 9.4 km). Further into *H. non-scripta* the remaining steep clines of 15 SNPs in eight loci presented similar shapes amongst each other (slope: range 0.057 – 0.081 km^{-1}, mean = 0.064 km^{-1}; δ_P: range 0.86 – 95, mean = 0.9; centre: range 16.15 – 24.06 km, mean = 19.44 km).

On the other side of the cline centre, for *H. hispanica*, the steep clines occurred more evenly distributed across the 1D transect. Towards the edge of *H. hispanica* steep clines with $\delta_P < 0.8$ were present, which was not observed for *H. non-scripta* (Figure 4.10). The allele frequencies for those variants revealed that *H. non-scripta* was mostly homozygous and that *H. hispanica* individuals were polymorphic for all three possible genotypes.

Gene functions of candidate genes. Taking the SNPs in these particular regions across the 1D transect and comparing them for overlapping or unique genes, the candidate genes for reproductive isolation were narrowed down to 32 different genes in unique positions across the 1D transect (Figure 4.11, Table 4.4). The gene with the steepest clines (in three different amplicons) overlapped in centre position with the morphology and organelle sections and was included nevertheless.

The gene identifications of these 32 genes were taken from *Musa acuminata* annotations of protein predictions (D'Hont et al., 2012) based on the methods outlined in chapter 3. Although the banana genome assembly excluded organellar genes and provided a chromosome map for the majority of genes (D'Hont et al., 2012), additional blast searches raised concerns regarding the origin of the matching bluebell references. The bluebell reference sequences were blasted against NCBI's non-redundant nucleotide database (BLASTn) and the protein swissprot database (BLASTx) and information was obtained from mostly gene predictions, conserved protein domains, and their linked GO-terms. In particular, genes could have been translocated from organelles into the nucleus (e.g. 'mitochondrial-like', 'chloroplastic-like') or be of organellar origin (e.g. 'mitochondrion', 'chloroplast'), albeit the latter annotation often referred to the target domain. The candidate genes with such annotations would henceforth need to be critically assessed in additional studies. Nonetheless, the banana annotations were used and key terms assigned (Table 4.4).

According to gene annotations, three genes have unknown functions, ten genes are related to function with the chloroplast and mitochondrion, three F-box genes that are involved in protein binding, three genes with carbohydrate metabolism, two genes identified as subunits of cellulose synthase A, five genes that either encode membrane proteins or are involved in transmembrane transport without further specification of biological pathways, and lastly seven genes linked to ion binding, transcription factor, serine/threonine

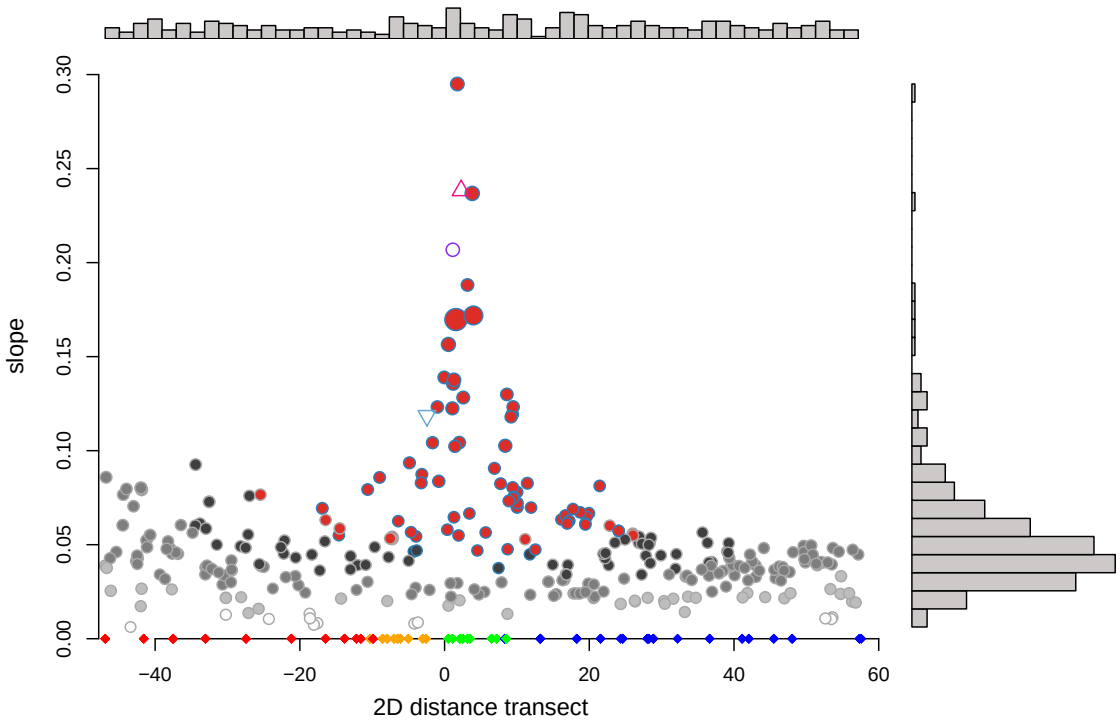


Figure 4.9 – Slope of a cline against its cline position for 316 biSNPs that have a significant slope and their centre position within the hybrid zone. Size of the dot represents the standard error of the slope (range $0.00173 - 0.07940$). Background of the dots is shaded by five categories of δ_P (white: 12 SNPs (0.16,0.33], light grey: 37 SNPs (0.33,0.5], grey: 128 SNPs (0.5,0.66], black: 59 SNPs (0.66,0.83], and red: 80 SNPs (0.83,1]). The 70 diagnostic loci are highlighted with blue circle margins. The histograms in the margin plots represent the number of SNPs in respective to their axis. Position and slope of the three alternative clines is given by hybrid index – purple circle, organelle cline – pink triangle, morphology cline – blue triangle. Filled diamonds at the bottom represent the position of the individuals along the 1D transect with blue – *H. non-scripta*, red – *H. hispanica*, green – hybrid North, orange – hybrid South.

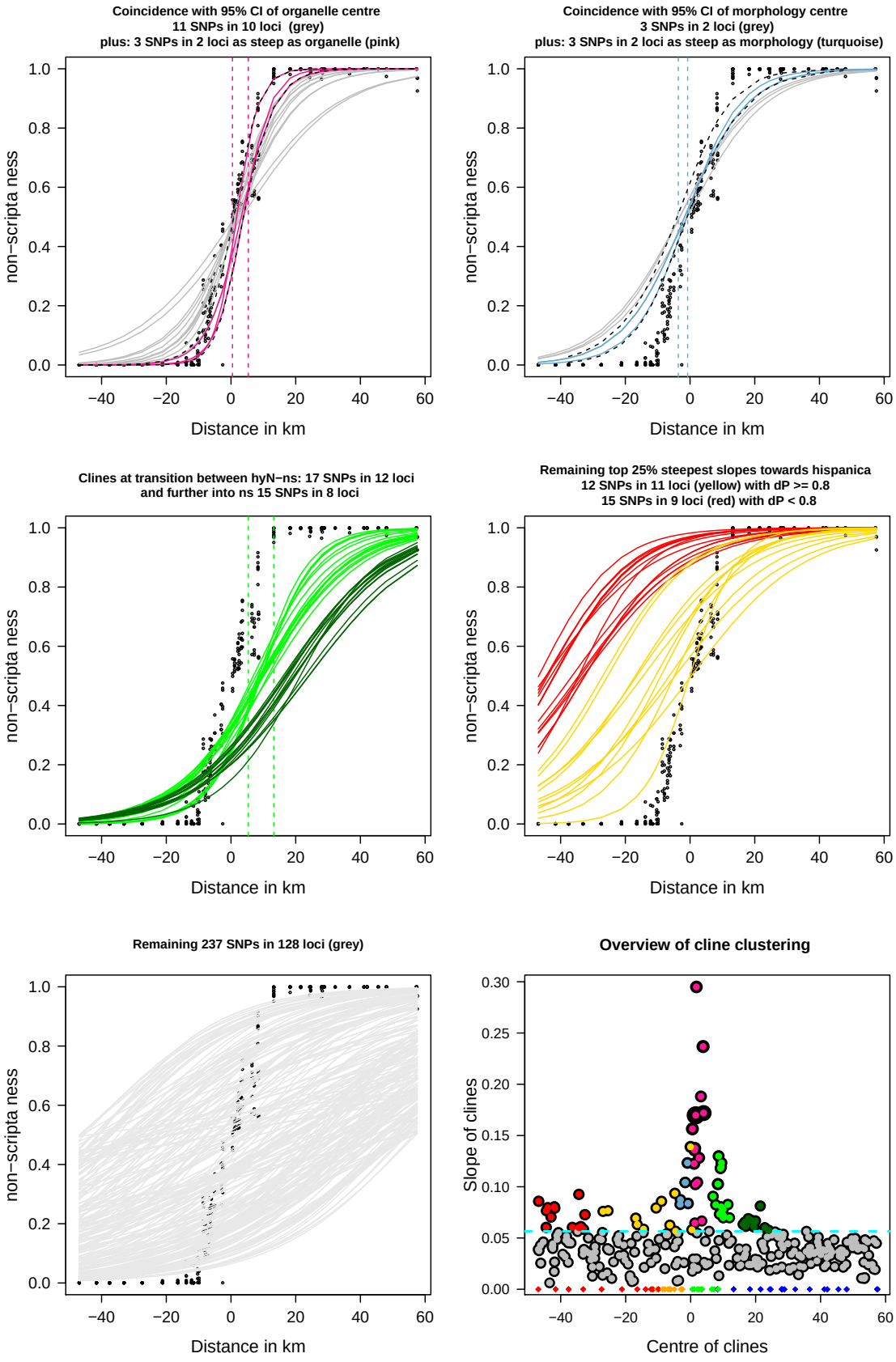


Figure 4.10 – Separating steepest top 25% clines by concordance with alternative markers and across the 1D transect. In bottom-right plot the blue dashed line represents the threshold of steep clines (slope $\geq 0.057\ km^{-1}$).

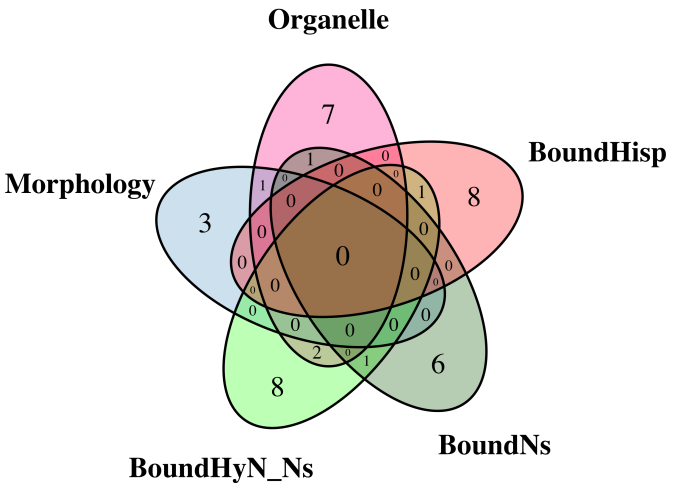


Figure 4.11 – Counts of unique genes for clines with top 25 % steep slopes and whose centres' coincide with certain regions across the 1D transect. The regions are: pink – 95% CI of organelle centre, lightblue – 95% CI of morphology centre, green – boundary between hybrid North and *H. non-scripta*, darkgreen – steep clines of *H. non-scripta*, red – region of *H. hispanica* with steep clines but $\delta_P < 0.8$.

kinase activity, ribosome biogenesis, DNA repair protein and immune response (Table 4.4). Two of the membrane genes also exhibited GO term annotation of cellular components of chloroplast and mitochondrion. The gene functions (key terms) cannot be clearly clustered by their cline centre positions (Table 4.4).

One gene, 'exocyst complex component 6' (GSMUA_AchrUn_randomG00330; 'protein complex involved in tethering vesicles to the plasma membrane during regulated or polarised secretion' – GO:0000145) with its centre position between hybrid North and *H. non-scripta* was mentioned to be involved in acceptance of pollen, pollen germination, and pollen tube growth in *Arabidopsis thaliana* (NCBI). But apart from that, none of the genes showed relevance for flower development.

Instead, one in three genes with strong genetic differentiation ($F_{ST} = 0.12 – 0.7$) and significant F_{ST} (between all four populations and total) can be linked to genetic processes relating to the organelles (e.g. photosynthesis such as RuBisCo, and membrane protein cytochrome C (*CytC*) of eukaryotic mitochondria). Five of those genes have their centre parameters coinciding with the boundary of *H. hispanica*, whereby polymorphisms were maintained in *H. hispanica* but *H. non-scripta* was mostly monomorphic. The folic acid synthesis protein *fol1* (mitochondrial in *Phoenix dactylifera*) included three amplicons totalling 17 SNPs and two of the amplicons exhibited three SNPs with steep clines caused by fixed allele frequency (AF) in *H. non-scripta* (mean AF = 0.97) and for the alternative allele in hybrid South (mean AF = 0.02) and *H. hispanica* (mean AF = 0.07), and an intermediate mean allele frequency AF = 0.56 of hybrid North (Figure 4.14).

Table 4.4 – Summary table of the 33 genes that showed top 25 % of steep slopes at a certain centre position across the 1D transect (Condition). The gene names were obtained from *M. acuminata*, and keys – used to characterise the gene function – were obtained from NCBI BLAST searches of the bluebell reference sequence. The cline parameters (slope, and centre) as well as the F_{ST} estimates were averaged (including standard deviation – SD) over all present significant SNPs (Cl) in the cline analyses, in contrast to the total number of discovered SNPs per gene (T).

		SNPs		Slope		Centre		Fst			
Condition	Bluebell reference	T	Cl	mean	SD	mean	SD	mean	SD	Gene annotation from *M. acuminata*	Key
BoundHisp	GSMUA_Achr4G23090|SWA2_36991	9	1	0.086		-46.781		0.124		Xylulose kinase	carbohydrates
BoundHisp	GSMUA_Achr5G06790|SWA2_280338	3	1	0.060		-44.473		0.090		expressed protein	membrane
BoundHisp	GSMUA_Achr9G05680|SWA2_278219	5	2	0.078	0.002	-44.184	0.315	0.144	0.004	MAC/Perforin domain containing protein, putative, expressed	immune response
BoundHisp	GSMUA_Achr11G05060|SWA2_542573	12	3	0.080	0.001	-41.928	0.062	0.151	0.003	Putative Pentatricopeptide repeat-containing protein At4g01990	mitochondrial like
BoundHisp	GSMUA_Achr10G28780|SWA2_289939	12	1	0.060		-36.438		0.171		zinc ion binding protein, putative, expressed	unknown; organelle
BoundHisp	GSMUA_Achr2G13970|SWA2_30359	9	1	0.093		-34.403		0.192		ATP synthase gamma chain 1, chloroplastic	chloroplastic-like
BoundHisp	GSMUA_Achr3G03330|SWA2_507734	14	2	0.061	0.001	-34.124	0.446	0.200	0.003	Molybdopterin biosynthesis protein CNX3	mitochondrial
BoundHisp	GSMUA_Achr6G25590|SWA2_543757	14	3	0.068	0.008	-32.685	0.255	0.225	0.017	Chlorophyllide a oxygenase, chloroplastic	chloroplastic
Morphology	GSMUA_Achr3G04190|SWA1_40808	12	2	0.104	0.000	-1.668	0.000	0.443	0.000	Cytochrome c oxidase assembly protein COX15 homolog	membrane: at mt/(cp?)
Morphology	GSMUA_Achr7G11390|SWA2_288979	3	1	0.123		-0.992		0.534		F-box/LRR-repeat protein 15	F-box
Morphology	GSMUA_Achr8G12790|SWA1_576349	12	1	0.084		-0.805		0.462		Putative expressed protein	unknown
Organelle	GSMUA_Achr11G22050|SWA2_473351	3	1	0.157		0.533		0.619		RuBisCO large subunit-binding protein subunit alpha, chloroplastic	chloroplastic
Organelle	GSMUA_Achr2G15610|SWA2_278439	14	1	0.123		1.074		0.540		Putative Serine/threonine-protein kinase HT1	chloroplast differentiation
Organelle	GSMUA_Achr8G16870|SWA2_546097	6	2	0.137	0.001	1.245	0.073	0.578	0.005	Putative DUF246 domain-containing protein At1g04910	membrane: at mt/(cp?)
Organelle	GSMUA_Achr1G18060|SWA2_41312	7	1	0.065		1.300		0.377		Putative Rhamnogalacturonate lyase	carbohydrates
Organelle	GSMUA_Achr1G07730|SWA1_82983	9	1	0.104		2.032		0.506		Putative 1-acylglycerophosphocholine O-acyltransferase 1	binding
Organelle	GSMUA_Achr8G28940|SWA2_48285	3	1	0.188		3.165		0.640		F-box/LRR-repeat protein 12	F-box
Organelle:Morphology	GSMUA_Achr9G09300|SWA2_43204	17	3	0.235	0.062	3.190	1.236	0.699	0.026	Putative Folic acid synthesis protein fol1	mitochondrion
Organelle	GSMUA_Achr5G06050|SWA2_16914	8	1	0.067		3.430		0.328		Probable cellulose synthase A catalytic subunit 1 [UDP-forming]	cellulose biosynthesis
BoundHyN_Ns	GSMUA_Achr8G24820|SWA2_457544	7	1	0.091		6.889		0.459		Putative Alkylated DNA repair protein alkB homolog 8	protection of DNA damage ?
BoundHyN_Ns	GSMUA_Achr9G06330|SWA1_51466	6	1	0.073		8.890		0.371		tesmin/TSO1-like CXC domain containing protein	transcription factor ?
BoundHyN_Ns	GSMUA_Achr9G06890|SWA2_66977	4	2	0.077	0.007	8.905	1.653	0.443	0.035	Probable receptor-like protein kinase At2g42960	kinase
BoundHyN_Ns	GSMUA_AchrUn_randomG09460|SWA1_322437	3	1	0.118		9.194		0.526		Probable cellulose synthase A catalytic subunit 2 [UDP-forming]	cellulose biosynthesis
BoundHyN_Ns	GSMUA_Achr7G22520|SWA2_51210	7	2	0.075	0.000	9.512	0.072	0.421	0.004	Putative Probable importin subunit beta-4	ribosome biogenesis ?
BoundHyN_Ns	GSMUA_Achr6G04590|SWA2_458449	7	1	0.078		10.025		0.438		Putative Subtilisin-like protease	immune response
BoundHyN_Ns	GSMUA_Achr5G01650|SWA2_457751	12	2	0.070	0.000	10.029	0.000	0.408	0.000	Heparanase-like protein 1	carbohydrates
BoundHyN_Ns	GSMUA_AchrUn_randomG00330|SWA2_545899	6	1	0.083		11.439		0.446		Probable exocyst complex component 6	membrane
BoundNs	GSMUA_Achr10G11820|SWA1_43326	18	4	0.064	0.001	16.818	0.506	0.382	0.010	Putative [Protein-PII] uridylyltransferase	binding
BoundNs	GSMUA_Achr6G18110|SWA2_276679	5	1	0.061		16.946		0.293		ubiquitin carboxyl-terminal hydrolase, family 1, putative	unkown
BoundNs	GSMUA_AchrUn_randomG25560|SWA2_275561	7	1	0.069		17.744		0.352		Fe(2+) transport protein 1	membrane transport
BoundNs	GSMUA_Achr4G28160|SWA1_534935	12	3	0.066	0.001	19.429	0.636	0.314	0.007	Bifunctional aspartokinase/homoserine dehydrogenase 1, chloroplastic	chloroplastic like
BoundNs	GSMUA_Achr6G28030|SWA2_269031	11	2	0.061	0.000	19.458	0.010	0.359	0.002	uncharacterized LOC105038787	unknown
BoundNs	GSMUA_Achr8G06320|SWA2_30223	6	2	0.057	0.000	24.058	0.000	0.290	0.000	Putative F-box protein At5g49610	F-box

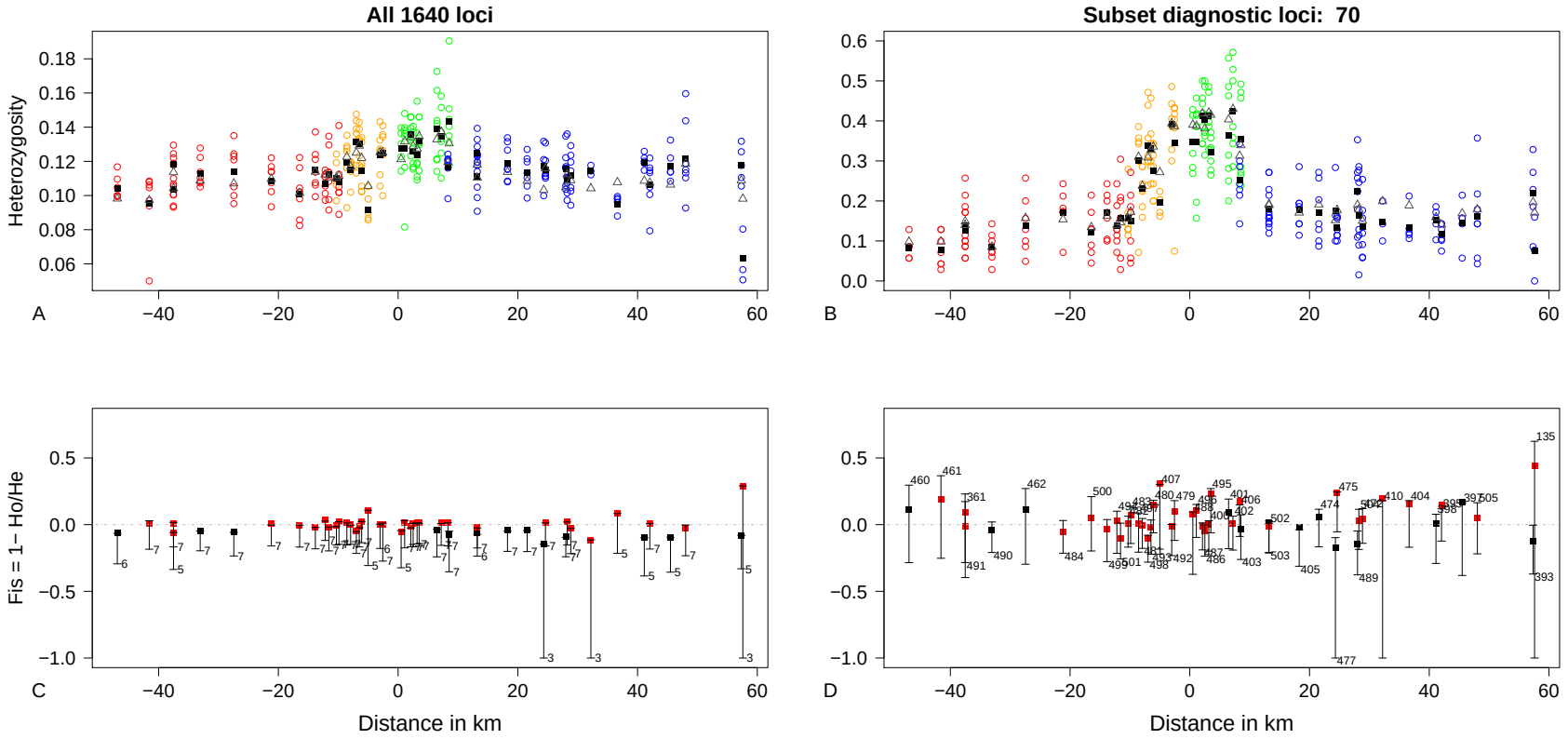


Figure 4.12 – First row represents individuals' observed heterozygosity (circles coloured with: red = *H. hispanica*, orange = hybrid South, green = hybrid North, and blue – *H. non-scripta*); along with their mean observed heterozygosity (black squares) and expected heterozygosity (grey triangles) per collecting site. Second row represents mean estimated inbreeding coefficient (F_{IS}) per collecting site with its 95 % confidence interval. If F_{IS} is significantly different from HWE squares are red. Left: full data set of 1,640 biSNPs. Right: subset of the 70 biSNPs with diagnostic allele frequencies.

4.3.4 Is there genetic evidence of heterosis-driven introgression?

Heterozygosity. The population genetic analyses showed that the hybrid individuals are admixed in a range of different proportions (Figure 4.6). Individuals of the northern hybrids showed more homogenised admixture proportions of 0.61 ± 0.076 compared to the southern hybrids with 0.22 ± 0.12. Also the AMOVA results showed no significant variation between individuals within sites for hybrid North. Heterozygosity was elevated across the population means of hybrids compared to the parent populations (Table 4.1), and looking at each collecting site across the 1D transect (expected and observed) heterozygosity was also increased in the hybrids – especially hybrid North (Figure 4.12 A). This pattern becomes even stronger for the subset of diagnostic alleles (Figure 4.12 B). However, at diagnostic loci the F_{IS} estimates per collecting site present a strong deficit of heterozygous genotypes (Figure 4.12 D). Nonetheless single loci can show genetic support for the hypothesis that heterozygote advantage might play a role in introgression by maintenance of alleles in the hybrid centre. Consequently, the cline analysis was repeated when removing either of the hybrid population to test this hypothesis at locus level.

Testing different slopes between the two hybrid populations. Performing simulations for all four populations and for removing either of the hybrid population, the bias on re-estimated slopes and centres was explored. In particular, the simulations estimated the expected allele frequency at a given collecting site across the 1D transect and randomly draw from it haplotypes for each individual at a given sites. Under this model, clines where achieved that predict the allele frequencies under independence between the two copies of the gene and random mating between samples within a collecting site. Further, a slope of 0.2068 was used as it corresponds to the results obtained from the hybrid index cline; and 0 for the intercept (see section 4.3.3). From 10^5 simulations 151 were discarded because they presented non-significant slopes of the cline in at least one of the three considered data sets.

The simulations showed very small shifts in slope and centre when removing either of the hybrid populations and comparing the resulting clines to the total set of four populations (All, Table 4.5). Removing one of the two central hybrid populations, the clines tend to be slightly sharper with larger standard errors (hyS and hyN, Table 4.5). However, the obtained distributions of cline and slope of hyS (i.e. hybrid North was removed) and hyN (i.e. hybrid South was removed) were not significantly different under this applied simulation scheme of panmixia and independent alleles (ks-test: $p = 0.978$; Table 4.5). Consequently, if we find some difference in the data, it is not just caused by removing a non-random part of the data but rather due to other evolutionary processes not considered by this model. In the simulated data, a general trend seems to be: when hybrid South is removed, the slopes are steeper with slightly larger standard errors, and shifted towards *H. non-scripta*. Lastly, to get an idea when the difference between two clines (in the observed data) may be significant, the distribution of differences of slope and centre between the two hybrid populations and their 95% confidence intervals were determined (Table 4.6).

The hypothesis was made that heterosis-driven introgression in more admixed northern hybrids (results 4.3.3, Figure 4.6) would be shown by shallower slopes due to more frequent

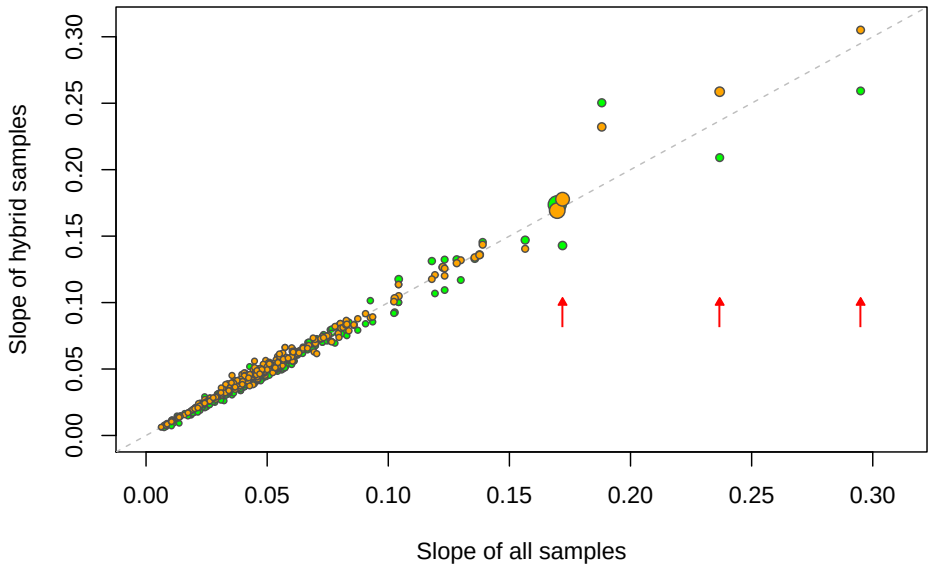


Figure 4.13 – The slope parameter (from repeated cline analysis) of hybrid North (green) and hybrid South (orange) are plotted against the slope for all samples (x-axis) to present the difference between hybrid slopes. The arrows mark the three biSNPs of significant difference between hybrid North and hybrid South. The size of the points are scaled by the standard error of the slope of the fitted cline.

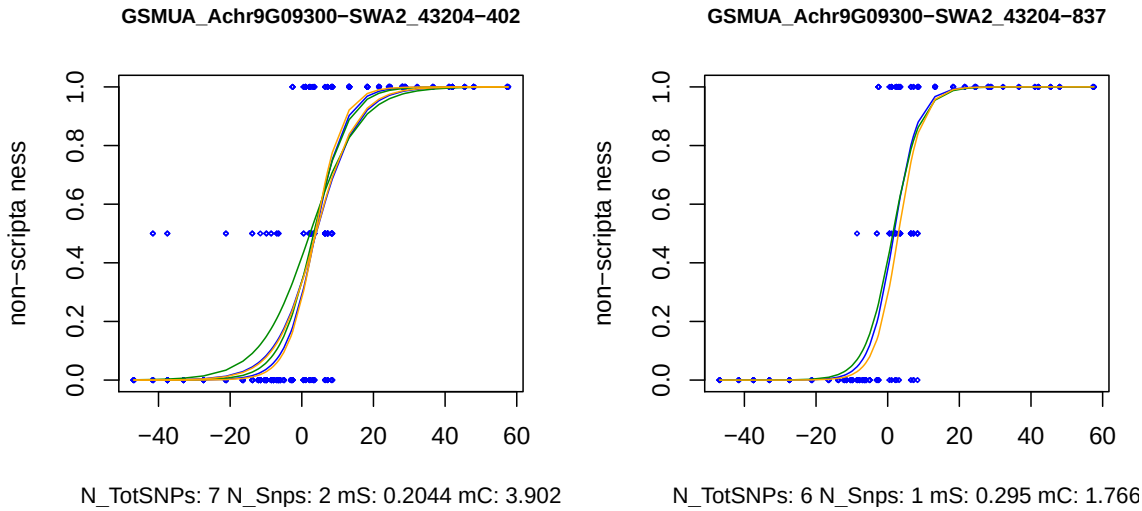


Figure 4.14 – Clines of the three steepest biSNPs in two different amplicons of the gene *fol1*. For each SNP the clines are shown for different numbers of samples: all samples – blue, hybrid North – green, hybrid South – orange. Blue diamonds mark the genotype by individual (0, 1, 2). The average cline parameters (slope – mS, centre – mC) are shown for the significant cline SNPs (N_SNPs) amongst all SNPs per gene (N_TotSNPs).

Table 4.5 – Mean estimate of the slope and its standard error (stdErr) and centre for the clines of three subsets of data and their corresponding pairwise Kolmogorov-Smirnov test results.

	All	hyS	hyN
Mean slope	0.2095	0.2101	0.2102
Mean stdErr	0.01737	0.01906	0.0194
All		$< 2.2e-16$	$< 2.2e-16$
hyS			0.978
Mean centre	-0.00269	-0.00268	-0.00251
All		$< 2.2e-16$	$< 2.2e-16$
hyS			0.7068

Table 4.6 – Mean slope and centre differences between the three different subsets and the 95 % confidence interval of the estimates. One-sample t-test was performed to test the mean difference in slope estimates for significant difference from zero.

	All vs no hyN	All vs no hyS	no hyN vs no hyS
Mean δ slope	-0.0006414	-0.0006726	-0.00003121
p-value	$< 2.2e-16$	$< 2.2e-16$	0.5257
2.5%	-0.02266	-0.02281	-0.03087
97.5%	0.01582	0.01608	0.03097
Mean δ centre	-0.00001507	-0.0001844	-0.0001693
p-value	0.9879	0.857	0.9152
2.5%	-0.6112	-0.6315	-0.9803
97.5%	0.6204	0.6354	0.9821

heterozygous individuals (GT = 1) and the standard errors of the slope would be larger because the residuals to the fitted clines are larger, in contrast for less admixed hybrid South the clines would be steeper (GT = 0 or GT = 2) and show smaller standard errors.

Therefore, applying the 95 % CI threshold of significantly larger slopes when removing hybrid North only one locus with three SNPs was found that fits this criterion. This gene, *fol1* (GSMUA_Achr9G09300|SWA2_43204), produced the steepest clines overall, coincided with the organelle cline and could be linked to function at the mitochondrion (section 4.3.3). The SNP's clines are shallower for the subset of northern hybrids than for the southern hybrids (Figure 4.13 and 4.14). However, the standard errors of the slopes were also smaller for hybrid North (0.0225 – 0.0229) than for hybrid South (0.0227 – 0.0775), which can be explained by the balanced genotype frequencies in hybrid North of exactly one third for each genotype (0, 1, 2) of the two biSNPs in the first amplicon, and an excess of *H. non-scripta* genotype (0) at the third biSNP in the third amplicon (0:0.51, 1:0.34, 2:0.15). So this locus alone might not present a strong argument for the maintenance of a cline through heterosis and that both alleles move forward simultaneously by introgression.

4.4 Discussion

4.4.1 Introduction

Hybrid zones are 'natural laboratories' (Hewitt, 1988), which can be used to explore the genetics of reproductive isolation on the one hand, and the amount of introgression on the other (e.g. Arnold and Martin, 2009; Barton and Hewitt, 1989; Nolte et al., 2009; Via, 2012; Walsh et al., 2016). Building on the results from chapter 2, which showed high hybrid fecundity, low reproductive isolation between the parental species, and a significant but small asymmetry in F_1 hybrid formation, the population dynamics will be discussed in the light of spatio-temporal dynamics of the hybrid zone between two closely related species, *Hyacinthoides non-scripta* and *H. hispanica*. In this study, cline analyses by generalised linear models per locus were used to identify genetic barriers to gene flow and to explore the potential for heterosis-driven introgression.

4.4.2 Marker bias and genetic diversity

The marker design was intended to enrich for diagnostic loci in expressed genes, which show nearly fixed allele frequencies in either parental species so that they segregate in their hybrids (chapter 3). Additionally, potentially conserved protein structures were favoured in the process because the bluebell mRNA were compared to a distant relative (*Musa acuminata*). Due to this design a substantial bias towards elevated genetic differentiation (F_{ST}) even in neutral alleles was expected. Our simulations exploring the marker bias showed a significant shift of F_{ST} for neutral markers, which was far from sufficient to explain the much stronger differentiation observed at the target sites. Consequently, the data set successfully captured biological patterns across the hybrid zone. The re-sequencing data presented low abundance of fixed markers (3.4 %) in a subset of parental samples across their distribution range (chapter 3), and across the parental samples of the hybrid zone (6.1 %). Possibly to our advantage, the lack of many fixed markers enabled us to observe a large amount of introgression in potentially neutral genes, and nonetheless, the large number of private alleles (especially singletons) were sufficient to differentiate between individuals, collecting sites, and populations.

Small F_{ST} between both parental populations ($F_{ST} = 0.083$) suggested very limited genetic differentiation at genic regions. This can be explained by local substructure within populations between collecting sites (e.g. isolation-by-distance test) and by genetic diversity within individuals. The former was shown by positive F_{IS} with significant deviations from HWE of the allele frequencies and the latter by highest contribution of individual variance components in AMOVA (> 74 %).

4.4.3 Shared polymorphisms as ancestral diversity

A larger proportion of shared polymorphisms with intermediate frequencies between *H. non-scripta* and *H. hispanica* was discovered than expected (Figure 4.7). The genetic markers in coding genes are under stronger selection pressures against mutations and genetic drift (balancing selection and codon bias, Hershberg and Petrov, 2008). Accordingly,

the shared polymorphisms could be caused by conserved markers under balancing selection to maintain crucial plant metabolisms and immune response (e.g. MHC in mammals, Charlesworth, 2006). However, under neutral expectations such trans-specific polymorphisms are unlikely to persist after speciation over an evolutionary time-scale unless they are maintained by gene flow (Charlesworth, 2006). According to the ABC results the shared polymorphisms could be ancestral genetic variation due to continuous but very small amounts of migration ever since an ancestral population split into *H. non-scripta* and *H. hispanica* about 1.15 – 3.69 mya (depending on the assumed generation time; section 4.7.2). The ABC results favoured a parapatric speciation model with ongoing gene flow, which has recently received more attention as a speciation model (Harrison and Larson, 2016; Nosil, 2008; Papadopulos et al., 2011). However, the ABC analysis tested only rather simplified coalescent histories of a maximum of two different populations. Given the local genetic diversity of bluebells in northern Spain the suggested models between only two populations might be rather unrealistic. Consequently, shared ancestral alleles and incomplete lineage sorting alone would not explain the gradual transition of alleles across the 1D transect (De La Torre et al., 2015).

Overall, the population structure without sympatric parental populations (based on nuclear biSNPs), the spread of both *H. non-scripta* and *H. hispanica* organelle haplotypes to the edges of hybrid distribution, and the hybrid index suggest a rather old hybrid zone with multiple generations of hybrids and backcrosses (Hamilton et al., 2013; Walsh et al., 2016). In addition, SNPs of different amplicons but from the same gene that showed slight differences in cline parameters suggest low linkage and effective recombination within targeted genes and further support the presence of an advanced generation hybrid zone (Walsh et al., 2016).

The complex geographic structure and changing climatic history during the Pleistocene (i.e. potentially ever since the inter-specific split) could have caused local survivals of either parental species (and old hybrids) in the Duero-Galician Mountains. The refugia-within-refugia hypothesis has been supported for several taxa in the Pyrenees and Cantabrian Mountains (Gómez and Lunt, 2006).

4.4.4 Shared polymorphisms due to contemporary introgressive hybridisation

The shared polymorphisms between the two species could be caused by extensive contemporary introgressive hybridisation, or be present as ancestral polymorphisms (e.g. van Herwerden et al., 2006). One way to distinguish between ancestral polymorphism and contemporary introgression is to examine the direction of the derived allele in cline analyses. A derived allele fixed in one population and gradually introgressing into the other could be explained by contemporary introgression, in contrast ancestral alleles would mostly be affected by drift and not accumulate a gradual pattern. Constant low gene flow would maintain the ancestral diversity through incomplete lineage sorting in a random geographic distribution of ancestral alleles (Edwards et al., 2008) and would consequently be absent from the geographic cline analyses. However, to determine in biSNPs which allele is the derived allele (i.e. new mutation) an outside group is required. Additionally, the SNP

data is limited because haplotype phasing has failed so far. Instead, the genotype data were transformed so that the most common allele always belongs to the *H. non-scripta* population, which forces the slope parameter to be always positive. Consequently, the direction of a cline (slope parameter is positive or negative) moving across the 1D transect cannot be assessed. Nonetheless, evidence in favour of contemporary introgression was found by the large number of shallow clines (237 biSNPs in 128 genes) that showed a significant transition of allele frequencies (width $\geq$ 17.5 km) across the 1D transect because they presented varying cline centres. Further, high inter-crossability evidenced by seed set and germination rate (chapter 2) showed a low reproductive barrier (e.g. Vega et al., 2013). Hybrid North represents a population of higher diversity and heterozygosity as well as admixture proportions of about 0.61, which is more indicative of early generation hybrids. In addition, it appears rather isolated from backcrosses with *H. non-scripta*, although no obvious geographic barrier is present. In contrast, hybrid South presents a more gradually transition to *H. hispanica* (low genetic differentiation), and the admixture proportions are lower with about 0.22, which is more representative of backcrosses with *H. hispanica* and to lesser extent with *H. non-scripta* or hybrid North along its valley. Such patterns have been found in other hybrid zone studies of recombinant hybrids (Christe et al., 2016; Lindtke et al., 2012). However, the possibility of backcrosses (that would facilitate introgression) was not assessed so far in bluebells by, for instance, cross-pollination experiments. Reports of invading *H. hispanica* (and a cultivar of *H. hispanica*) into several regions of *H. non-scripta*'s distribution range, probably mainly dispersed by humans since the 17^{th} century, and frequent reports of their hybrids evidence that contemporary hybridisation is occurring in natural habitats elsewhere (Grundmann et al., 2010; Kohn et al., 2009; Pilgrim and Hutchinson, 2004; Preston et al., 2002).

4.4.5 Differential introgression of organellar and nuclear markers

The geographic centre of the hybrid zone based on nuclear markers was found west and north of the Sierra del Teleno and does not coincide with its mountain peaks, which could have presented a population trough and reproductive barrier due to restricted pollen flow and seed dispersal between hybrid North and hybrid South as previously suggested (chapter 2). The mean of cline centres with low slopes was shifted towards *H. non-scripta* beyond the centre of the hybrid index, which would be indicative of *H. hispanica* alleles introgressing into *H. non-scripta*. Conversely, the alternative hypothesis that *H. non-scripta* introgressed into *H. hispanica* seems more likely given the following evidence: The organelle cline centre was shifted towards *H. non-scripta* (Table 4.3). Indeed, the *H. hispanica* organelle haplotype dominated in hybrid samples (mean 0.65) and reached up north close to *H. non-scripta*'s collecting sites. The organelle marker represents the distribution of mother genotypes because as commonly found in flowering plants (Birky, 2008), plastids are also maternally inherited in bluebells by seeds (Sears, 1980). In addition, bluebells present a very slow dispersal rate (0.6 – 6 cm/year, van der Veken et al., 2007), and their seeds show no adaptation to effective seed dispersal (Knight, 1964). The slow movement of organelle alleles across the hybrid zone is also shown by the very steep slope of the organelle cline. Consequently, bulbs of *H. hispanica* must have been distributed further north and were swamped by *H. non-scripta* pollen flow (Arrigo et al., 2011). It seems

likely that differential introgression is present between the nuclear (mostly free introgression) and organelle (heavily restricted by seed dispersal) genetic markers. The genes under selection would introgress fastest into *H. hispanica* (red and yellow clines in Figure 4.10), followed by neutral ones, and lastly genes trail behind that are linked to the organelles (green clines in Figure 4.10) in a way of cyto-nuclear incompatibilities (section 4.4.9). A substantial amount of gene flow by pollen would be required for this scenario.

Differential introgression can be caused by differential adaptation of parents to a changing climate and shifting the species boundaries in general, gradients of population density, and 'dominance drive' displacing recessive allele (e.g. genetic incompatibilities) (Buggs, 2007).

4.4.6 Asymmetric hybridisation

Asymmetric hybridisation can also explain the shift in nuclear biSNPs (Edwards et al., 2008). Based on genetic differentiation (F_{ST}), admixture proportions, and the principal component analysis, a stronger separation between hybrid North and *H. non-scripta* was shown than between hybrid South and *H. hispanica*. The pattern could be explained by asymmetric hybridisation of *H. non-scripta* into *H. hispanica*. On the one hand, the lower differentiation between hybrid South and *H. hispanica* can be caused by frequent backcrosses, which homogenise the genetic differentiation. The PCA visualised the genetic mixture of both populations clearly (Figure 4.2). The crossing experiments presented a slight advantage in F_1 formation if *H. hispanica* was the ovule donor. F_1 hybrids are, however, unlikely for both hybrid populations because in either region pure parental plants are absent in sufficient distance of pollen flow and seed dispersal, especially in order to explain the *H. hispanica* organelle haplotypes amongst northern hybrids.

On the other hand, steep clines (in eight different genes) were found at the boundary between hybrid North and *H. non-scripta*, which evidence restricted introgression due to a lack of backcrosses with *H. non-scripta*. There are no obvious physical barriers (i.e. altitude in mountain peaks) to separate these two populations, and gene annotations provide no accumulation of genes with cytoplasm-nuclear interactions in that region. It might be possible that morphological variation in flower shape contributes to the separation (see discussion in chapter 2). However, none of the clinal genes in bluebells are related to flower development and the morphological cline was shallow. The cline centre of morphological transition was positioned between hybrid North and hybrid South and coincided with the mountain peaks from Montes Aquilianos to the Sierra del Teleno that could pose a barrier to pollen flow. Overall, however, morphological variation seems to freely float (with pollen dispersal) across the hybrid zone. Another explanation of the lack of intermediate hybrids between *H. non-scripta* and hybrid North could be that they were missed in field collection west of the Montes Aquilianos and near BB-406, which was identified as *H. non-scripta* but showed about 13 % of admixture from *H. hispanica*.

Another barrier to gene flow could be flowering time, which has been discussed in chapter 2. If *H. hispanica* flowers later than plants northern of the hybrid zone, it would fit the pattern of asymmetric hybridisation. *H. non-scripta* would be able to disperse its pollen further into the hybrid zone than *H. hispanica*. Consequently, a pattern with hybrid North receiving pollen from *H. non-scripta* but less so reversed, and hybrid South receiving

pollen from hybrid North and rarely from *H. non-scripta* (exception BB-406), and more pollen into *H. hispanica* from hybrid South would support the observed admixture between *H. hispanica* and hybrid South.

4.4.7 Movement of species ranges and the hybrid zone due to climate change

Differential introgression between markers and asymmetrical hybridisation between parental taxa could present evidence of a moving hybrid zone (Buggs, 2007). Especially so in north-west Iberian Peninsula and the Cantabrian Mountains, which have been strongly impacted by the Quaternary climatic oscillations (Iriarte-Chiapusso et al., 2016; Schmitt, 2007; Serrano et al., 2016; Tarroso et al., 2016; Yustos and Martín, 2015) and driven movement of species (Iriarte-Chiapusso et al., 2016; Ramil-Rego et al., 1998; Remon et al., 2013; Taberlet et al., 1998). At the end of the last glacial cycle the main limitations to woodlands in north-west Iberia were low temperatures and dryness, with the most demanding tree species persisting on the seawards slopes of the Cantabrian Mountains (Iriarte-Chiapusso et al., 2016). While, the higher and inner mountain ranges, such as the Duero-Galician Mountains were suggested as refugia for pines during the Late-Glacial and Early Holocene (Iriarte-Chiapusso et al., 2016, and references therein). Based on the association of bluebells with deciduous forest dominated by *Quercus* (chapter 2; Blackman and Rutter, 1954) bluebells likely followed the re-colonisation routes of oak species, which supposedly recolonised the Galician-Duero Mountains ever since the Early Holocene (11,700 – 8,500 cal B.P.) from the southern and western lowlands (Brewer et al., 2002; Iriarte-Chiapusso et al., 2016). Bluebells can also dominate open uplands such as heath and grassland area as late successional ecosystems on deforested regions in Wales, UK (Ebuele et al., 2016). Possibly, the Buerzo basin and river Síl might have provided a nearby lowland sanctuary for bluebells during the colder phases (e.g. the Younger Dryas 17,000 cal B.P.; 11.2 ka event; 8.2 ka event; Iriarte-Chiapusso et al., 2016; Sobrino et al., 2007), which were partially inhabited by deciduous forests. Given that the climate has been cooling (-0.3°C/ky, Wanner et al., 2015) since the Holocene thermal maximum (ca 8,000 B.P., Ljungqvist, 2011) a southwards movement of *H. non-scripta* caused by climate change could be possible. However, proving a hybrid zone movement is challenging due to its conflicting causes and multiple explanations for patterns in molecular data (Buggs, 2007).

4.4.8 Heterosis-driven introgression

The pattern of intermediate hybrid index and strongly elevated heterozygosity at diagnostic loci and throughout all markers in hybrid North remains interesting. High levels of heterozygosity and hybrid advantage are often expected for F_1 hybrids (Anderson and Thompson, 2002; Lippman and Zamir, 2007). This phenomenon is referred to as heterosis or hybrid vigour (Shull, 1908), which is the 'phenotypic superiority of a hybrid over its parents' (Lippman and Zamir, 2007). Heterosis in plants is a multigenic complex, which can affect the phenotypic level in varying traits such as vegetative growth, flowering time, and seed yield (Lippman and Zamir, 2007). However, phenotypic superiority

in itself does not need to result in higher reproductive fitness. The genetic mechanism implies that hybrids experience a fitness advantage (reproductive success) because their heterozygous alleles obtained from either parent complement (slightly) deleterious recessive alleles and provide genetic plasticity to environmental conditions (Baranwal et al., 2012; Lippman and Zamir, 2007). As underlying genetic models, dominance (including over-dominance, and pseudo-overdominance), epistatic interactions and epigenetic factors have all been suggested to explain heterosis (Baranwal et al., 2012). Heterosis is mostly present as genome-wide heterozygosity in F_1 hybrids, which can be quickly lost by sexual reproduction and through recombination (breaking up epistasis), and subsequent hybrid generations experience so-called 'hybrid breakdown' (Lindtke et al., 2012; Lippman and Zamir, 2007). In contrast, fitness advantage caused by heterozygosity at specific loci can persist in a population despite recombination (Lippman and Zamir, 2007). Such alleles could spread and homogenise the population quickly by adaptive introgression (Nolte et al., 2009), which would present a pattern of lower genetic differentiation (e.g. F_{ST}) and shallower cline slopes. Accordingly, the hypothesis was formulated that heterosis for specific loci could be driving partial introgression. It was tested by repeated cline analyses that contrasted both hybrid populations. As a result, the SNPs with significantly different cline slopes actually presented a lack of heterozygous alleles in the northern hybrid population. Accordingly, the results from this cline approach were not satisfying the hypothesis. In our approach only single locus-specific heterozygosity was examined, but not the interaction between multiple loci and we might have failed to capture the respective loci with advantageous phenotypic traits. Alternatively, outlier scans of F_{IS} loci could be performed, especially, when assuming that loci under balancing selection would introgress quickly across the complete hybrid zone and not just within either hybrid population.

However, heterosis-driven partial introgression would have contradicted the observation of potential cyto-nuclear incompatibilities that reduce heterozygosity at loci (see discussion below).

4.4.9 Cyto-nuclear interactions as reproductive barrier

In the face of gene flow it is possible to accumulate substantial genetic differences due to ecological adaptions and consequently divergent selection (Nosil, 2008; Papadopulos et al., 2011; Via, 2012). Between the two parental species 45 different genes that presented diagnostic markers (i.e. markers that are highly differentiated) were found. A number of candidate genes for promoting differentiation between both parental populations was shown based on steep clines and their unique occurrence across the 1D transect.

Amongst the steepest clines were genes that are involved in photosynthesis (e.g. RuBisCo binding protein), transmembrane transports between cellular components, and biosynthesis processes that involve plastids and/or mitochondria. In particular, the gene putatively encoding for folic acid synthesis protein (*fol1*) coincided with the organelle cline and was even steeper (average width 5.7 km). Based on a lack of heterozygous allele frequencies in the hybrid individuals at the *fol1* gene, there seems to be strong divergent selection on the alleles in relation to the samples' organelle haplotype. Folic acid, also termed vitamin B9 in human dietary supplements (Lucock, 2000), is part of the folate biosynthesis process that transports and donates C1-units, which are in plants usually

methyl (or its redox form) groups (Rebeille et al., 2007). The biosynthesis of folate is organised in three different compartments, the cytosol, the plastid, and the mitochondria, in which crucial steps of the folate synthesis are made (Rebeille et al., 2007; Sahr et al., 2005). Also other genes, coinciding with the organelle cline, present interactions between different cellular components (Table 4.4).

The plant metabolism and other cell functions are strongly compartmentalised between different parts of the cells (Lunn, 2007). Many of the genes that facilitate the function of photosynthesis and carbohydrate metabolism have been transferred from the organellar genomes into the nuclear genome (Adams and Palmer, 2003; Greiner and Bock, 2013; Martin and Herrmann, 1998). Therefore, many proteins and enzymes that regulate functioning of mitochondria and plastids are translated from nuclear genes and need to be transported to their target domain, which demonstrates the tight coordination and co-evolution of nuclear and organelle functions (Greiner and Bock, 2013). In most plants the organelles are maternally inherited and lack recombination due to asexual transmission (Birky, 2008), in contrast to the nuclear genome in sexually reproducing species (Greiner and Bock, 2013).

Such cyto-nuclear interactions can lead to so-called intrinsic Bateson-Dobzhansky-Muller incompatibilities (BDMI) (Bateson, 1909; Dobzhansky, 1937; Muller, 1942). The model assumes that incompatibilities arise when two ancestral populations are isolated and accumulate neutral mutations due to drift, which, when coming back into contact due to hybridisation, have not been tested together and lead to less fit hybrids. Often, cyto-nuclear interactions respond to divergent ecological adaptation because they affect photosynthesis and respiratory metabolisms (Burton et al., 2013; Lunn, 2007). For instance, artificial hybrids between *Atropa belladonna* and *Nicotiana tabacum* showed reduced photosynthetic ability due to a failure of nuclear-encoded RNA editing nucleotides in the *atpA* gene of the tobacco plastome (see Burton et al., 2013). Given that one in three genes with potential for reproductive isolations indicate possible interactions with the organelle genome, and of these four genes coincide with the transition of organelle haplotypes, this can hint towards BDMI in this bluebell hybrid zone.

At this point, the presence of BDMI contradict previous observations in that they generally cause lower hybrid fitness compared to the parents, and that they become effective in secondary contact. BDMI lower the fitness of hybrids, although this can range from complete failure of reproduction, to cytoplasmic male sterility (Greiner and Bock, 2013), and a range of phenotypic changes as found for *Helianthus* hybrids (Levin, 2003; Sambatti et al., 2008). In bluebells, inter-specific and hybrid crosses showed a higher seed set for outcrosses, however, they do not allow assumptions about long-term fitness of F_1 and later generation hybrids. A BDMI could be recessive and thus only expressed in the second generation cross, i.e. if hybrids are homozygous at the BDMI locus, the hybrid seed crosses would fail to form the heterozygote formation that produces lower fitness and would also only be present in the second hybrid generation between hybrids when the recessive alleles are homozygous again.

BDMI affect the post-zygotic fitness, for instance photosynthesis efficiency or carbon metabolism, it could become effective in the first year of seedling growth, when the seeds resources are exhausted. The lack of deficient hybrids in the field ('hybrid breakdown',

i.e. reduced fitness compared to the parental lineage, Burton et al., 2013) so far could be explained by the long bluebell life-cycle. The leaves are rather inconspicuous and flowers only appear after five years (Woodhead, 1904), or when the bulbs reach a dry mass of 0.2 - 0.4 g (Blackman and Rutter, 1954). Consequently, a hybrid breakdown could inhibit unsuitable genotype combinations reaching reproductive age and such plants were missed in the field work. Further crossing experiments including backcrosses and long term growth assessments of the F_1 and F_2 hybrids focusing on the recombination of divergent nuclear and organelle genotypes could further test the presence of BDMI (e.g. Sambatti et al., 2008).

Lastly, BDMI require sufficient divergence between alleles to facilitate incompatible interactions, which can be achieved by allopatric isolation and by divergent local adaptation in parapatry. The genetic differentiation between organelles in bluebells is relatively high (19 SNPs in 13 genes) with a low intra-specific genetic variation. Instead, at the nuclear level the genetic differentiation was low and parapatric speciation was suggested. If the organelle haplotype were locally adapted to a specific ecological environment, coevolving nuclear loci could subsequently present adaptive divergence to the environment as well (Arnold and Martin, 2009; Burton et al., 2013). For example, the importance of cyto-nuclear incompatibilities in reproductive isolation was exemplified in different sunflower species by replacing the cytoplasm of *Helianthus annuus* by cytoplasm of conspecifics (Levin, 2003). And further, transplant experiments of backcrosses between two sunflower species, *Helianthus annuus* and *H. petiolaris*, showed that organelle haplotype might influence the ability to cope with drought stress (reduced fitness in hybrids and asymmetrical in backcrosses) and that they drive ecological differentiation due to cyto-nuclear incompatibilities (Sambatti et al., 2008). In this particular north-western Iberian region, the transition from Atlantic climate to Mediterranean climate (Eurosiberian/Mediterranean border, Sobrino et al., 2007) along the Cantabrian Mountains and the Duero-Galician mountains might pose an ecological boundary to introgressing *H. non-scripta* alleles.

Another problem remains, especially for loci showing strong differentiation between parental populations, to differentiate between alleles that originated from allopatric speciation, and those alleles that diverged more recently due to e.g. ecological adaptation across the hybrid zone (Harrison and Larson, 2016). Employing additional analyses, such as NEWHYBRIDS (Anderson and Thompson, 2002), that identify potential generation-class of hybrid individuals based on similar Bayesian co-ancestry such as STRUCTURE would provide additional insights into the age structure of the bluebell hybrid zone (e.g. Vähä and Primmer, 2006; Vega et al., 2013). Additionally, a larger and random genome-wide SNP data set, regardless of coding or non-coding region, would be more suitable to test ancestral and recent coalescent histories between *H. non-scripta* and *H. hispanica* (e.g. Christe et al., 2016; Harrison and Larson, 2016). The challenge will be to obtain such data for non-model species with large genome sizes, as previously discussed in chapter 3.

4.5 Conclusion

The genetic differentiation (F_{ST}) between *H. non-scripta* and *H. hispanica* was found to be relatively low in the hybrid zone, possibly due to a long-term interaction by pollen

flow, maintenance of local genetic diversity, and shifting distribution ranges throughout Holocene climate oscillations. The large amount of shared polymorphisms could either be due to ancestral shared polymorphisms (of conserved genes) or contemporary introgression at neutral alleles. Under the assumption of contemporary introgression that occurs across both hybrid populations, a possible southwards movement of the hybrid zone was suggested, which could be caused by asymmetric hybridisation, differential introgression and climate change. Indeed, hybrid North seems isolated from backcrosses with both parents but with possible gene flow from *H. non-scripta*. Hybrid South presents the possibility of frequent backcrosses with *H. hispanica* due to their (genetic and spatial) proximity within the valley. Pollen influx of *H. non-scripta* could be possible from localities such as BB-406, or from hybrid North.

Genetic differentiation at the organelle genomes must have occurred during a period of restricted gene flow because the current set of individuals shows cytoplasmic-nuclear incompatibilities that select against heterozygosity in the centre of the zone. Therefore, post-zygotic hybrid breakdown must have been overlooked so far, or are only expressed later, either in life cycle or in F_2 hybrids. Ecological adaptations based on organellar haplotype could drive adaptations to divergent selection in nuclear genes due to the cyto-nuclear incompatibilities. Consequently, divergence could be accumulated despite gene flow and facilitate parapatric speciation. Recently, Höllinger and Hermisson (In prep.) determined under which conditions cyto-nuclear incompatibilities can indeed be maintained despite gene flow in a continent island model.

These results lay out hypotheses for a natural hybrid zone between *H. non-scripta* and *H. hispanica* that need further testing in future studies. Such studies could focus on resampling the hybrid zone at a later time point to examine a movement of the zone; further crossing and transplant experiments to confirm BDMIs and their potential for ecological divergence; ecological niche modelling of past distribution ranges could clarify the chance of parapatric speciation in contrast to secondary contact.

4.6 Contributions and acknowledgements

This project was part of a collaboration between Alexandre Blanckaert and myself within the EU-funded INTERCROSSING network. Alexandre was supervised by Prof. Joachim Hermisson at University of Vienna from the Faculty of Mathematics. He was particularly involved in the approximate Bayesian computations and the theoretical expectations of allelic distributions.

4.7 Supplement information

4.7.1 Supplementary tables

Table 4.7 – AMOVA table for three or four populations with the results of different levels of variability in the rows. The component of covariance (σ) and their contribution to the total covariance (%) are also reported. Significance of covariance was assessed using 1000 permutations with: $p \leq 0.05$ (*), $p \leq 0.01$ (**), and $p \leq 0.001$ (***).

Level of differentiation	Df	Sum Sq	Mean Sq	σ		%
Between Pop3	2	8680.05	4340.02	20.43		15.86
Between Sites Within Pop3	45	9372.06	208.27	7.95	***	6.17
Between samples Within Sites	259	27648.22	106.75	6.29	***	4.88
Within samples	307	28911.57	94.17	94.17	***	73.09
Total	613	74611.89	121.72	128.85		100.00
Between Pop4	3	9806.37	3268.79	20.44		16.06
Between Sites Within Pop4	44	8245.74	187.40	6.33	***	4.97
Between samples Within Sites	259	27648.22	106.75	6.29	***	4.94
Within samples	307	28911.57	94.17	94.17	***	74.02
Total	613	74611.89	121.72	127.23		100.00

Table 4.8 – Overview table of included collecting sites ordered by their distance to the 1D cline centre (dist1D). The populations (Pops) are abbreviated as hisp = *H. hispanica*, hyS = hybrid South, hyN = hybrid North, ns = *H. non-scripta*. N provides the number of samples. Longitude (Lon), latitude (Lat), and altitude in m above mean sea level (Alt) are also listed. The sites' mean admixture proportions for K=2 (Adm mean ± SD), mean chloroplast frequency (CPF mean ± SD), and inbreeding coefficient (F_{IS}; significance obtained by 1000 permutations of individuals' haplotype with: $p \leq 0.05$ (*), $p \leq 0.01$ (**), and $p \leq 0.001$ (***); 95% confidence intervals CI were obtained by bootstrapping) are given.

Site	N	Pops	Lon	Lat	Alt	dist1D	Adm mean ± SD	CPF mean ± SD	Fis		95% CI
460	6	hisp	-5.98	41.84	775.00	-46.90	0.00 ± 0.00	0.00 ± 0.00	-0.07	ns	(-0.29, -0.06)
461	7	hisp	-6.21	41.88	980.00	-41.57	0.00 ± 0.00	0.00 ± 0.00	0.01	***	(-0.18, 0.03)
491	7	hisp	-6.46	42.03	936.24	-37.53	0.00 ± 0.00	0.00 ± 0.00	0.01	***	(-0.17, 0.02)
361	5	hisp	-6.71	41.96	770.00	-37.51	0.00 ± 0.00	0.00 ± 0.00	-0.06	***	(-0.34, -0.06)
490	7	hisp	-6.41	42.05	947.49	-33.06	0.00 ± 0.00	0.00 ± 0.00	-0.05	ns	(-0.20, -0.05)
462	7	hisp	-6.32	42.06	853.37	-27.44	0.00 ± 0.00	0.00 ± 0.00	-0.05	ns	(-0.23, -0.06)
484	7	hisp	-6.73	42.11	1009.38	-21.18	0.00 ± 0.00	0.00 ± 0.00	0.01	***	(-0.16, -0.00)
500	7	hisp	-6.59	42.27	970.07	-16.45	0.00 ± 0.00	0.00 ± 0.00	-0.00	***	(-0.17, -0.00)
499	7	hisp	-6.63	42.26	984.65	-13.82	0.02 ± 0.03	0.00 ± 0.00	-0.02	***	(-0.18, -0.02)
494	7	hisp	-6.49	42.31	844.43	-12.18	0.01 ± 0.02	0.00 ± 0.00	0.04	***	(-0.12, 0.03)
501	7	hisp	-6.23	42.20	945.36	-11.57	0.01 ± 0.02	0.00 ± 0.00	-0.02	***	(-0.20, -0.02)
482	7	hyS	-6.61	42.35	930.19	-10.31	0.04 ± 0.03	0.00 ± 0.00	-0.01	***	(-0.15, -0.01)
483	7	hisp	-6.52	42.34	920.48	-9.88	0.04 ± 0.04	0.00 ± 0.00	0.02	***	(-0.15, 0.02)
497	7	hyS	-6.64	42.38	785.20	-8.56	0.21 ± 0.06	0.14 ± 0.38	0.01	***	(-0.15, 0.01)
481	7	hyS	-6.62	42.38	971.02	-7.93	0.12 ± 0.05	0.29 ± 0.49	0.00	***	(-0.15, -0.00)
498	7	hyS	-6.68	42.34	866.18	-6.97	0.15 ± 0.05	0.00 ± 0.00	-0.04	***	(-0.22, -0.05)
493	7	hyS	-6.62	42.40	1006.24	-6.49	0.22 ± 0.04	0.00 ± 0.00	-0.01	***	(-0.17, -0.02)
480	7	hyS	-6.63	42.40	855.54	-6.09	0.24 ± 0.05	0.14 ± 0.38	0.03	***	(-0.14, 0.02)
407	5	hyS	-6.66	42.42	560.12	-5.01	0.30 ± 0.05	0.00 ± 0.00	0.11	***	(-0.31, 0.11)
492	6	hyS	-6.69	42.41	684.25	-3.02	0.34 ± 0.06	0.50 ± 0.55	0.00	***	(-0.18, 0.00)
479	7	hyS	-6.70	42.41	672.42	-2.56	0.35 ± 0.17	0.57 ± 0.53	-0.00	***	(-0.27, 0.01)
488	5	hyN	-6.43	42.42	1230.22	0.51	0.49 ± 0.04	0.20 ± 0.45	-0.06	***	(-0.32, -0.04)
496	7	hyN	-6.55	42.45	780.21	1.09	0.54 ± 0.02	0.43 ± 0.53	0.01	***	(-0.17, 0.03)
487	7	hyN	-6.32	42.38	1068.78	2.12	0.57 ± 0.04	0.71 ± 0.49	-0.01	***	(-0.17, -0.02)
486	7	hyN	-6.14	42.32	910.42	2.51	0.58 ± 0.04	0.43 ± 0.53	0.01	***	(-0.17, 0.01)
400	7	hyN	-6.56	42.48	675.00	3.17	0.61 ± 0.04	0.29 ± 0.49	0.02	***	(-0.15, 0.01)
495	7	hyN	-6.53	42.46	1054.87	3.51	0.68 ± 0.06	0.57 ± 0.53	0.01	***	(-0.16, 0.01)
401	7	hyN	-6.52	42.50	963.33	6.48	0.68 ± 0.06	1.00 ± 0.00	-0.04	ns	(-0.24, -0.03)
402	7	hyN	-6.48	42.49	1148.19	7.20	0.63 ± 0.05	0.43 ± 0.53	0.01	***	(-0.16, 0.01)
406	7	ns	-6.82	42.46	719.72	8.33	0.87 ± 0.04	1.00 ± 0.00	0.02	***	(-0.16, 0.01)
403	7	hyN	-6.46	42.50	1079.00	8.53	0.64 ± 0.08	0.43 ± 0.53	-0.08	ns	(-0.35, -0.05)
502	6	ns	-6.49	42.55	770.71	13.22	0.99 ± 0.01	1.00 ± 0.00	-0.07	ns	(-0.24, -0.07)
503	7	ns	-6.46	42.54	758.56	13.22	0.98 ± 0.02	1.00 ± 0.00	-0.02	***	(-0.17, -0.02)
405	7	ns	-6.44	42.58	812.00	18.27	1.00 ± 0.01	1.00 ± 0.00	-0.04	ns	(-0.20, -0.05)
474	7	ns	-6.72	42.66	534.93	21.55	1.00 ± 0.00	1.00 ± 0.00	-0.04	ns	(-0.20, -0.04)
477	3	ns	-6.89	42.63	696.77	24.32	0.99 ± 0.02	1.00 ± 0.00	-0.15	ns	(-1.00, -0.15)
475	7	ns	-6.87	42.64	554.05	24.61	1.00 ± 0.01	1.00 ± 0.00	0.01	***	(-0.16, 0.01)
489	7	ns	-6.27	42.64	813.16	28.01	0.99 ± 0.01	1.00 ± 0.00	-0.09	ns	(-0.24, -0.10)
504	7	ns	-6.32	42.65	825.79	28.24	1.00 ± 0.01	1.00 ± 0.00	0.02	***	(-0.15, 0.02)
472	7	ns	-6.48	42.70	805.20	28.85	1.00 ± 0.00	1.00 ± 0.00	-0.03	***	(-0.22, -0.03)
410	3	ns	-7.46	42.28	1010.00	32.20	1.00 ± 0.00	1.00 ± 0.00	-0.11	***	(-1.00, -0.11)
404	5	ns	-6.22	42.71	950.00	36.63	1.00 ± 0.01	1.00 ± 0.00	0.08	***	(-0.22, 0.08)
398	5	ns	-7.10	42.69	1060.00	41.12	1.00 ± 0.00	1.00 ± 0.00	-0.10	ns	(-0.38, -0.10)
395	7	ns	-6.51	42.83	882.00	42.05	1.00 ± 0.00	1.00 ± 0.00	0.01	***	(-0.18, 0.01)
397	5	ns	-7.08	42.76	825.00	45.50	1.00 ± 0.00	1.00 ± 0.00	-0.09	ns	(-0.36, -0.09)
505	7	ns	-6.45	42.88	850.05	48.03	1.00 ± 0.01	1.00 ± 0.00	-0.03	***	(-0.23, -0.00)
393	5	ns	-6.23	42.91	1260.00	57.33	1.00 ± 0.00	1.00 ± 0.00	-0.08	ns	(-0.33, -0.08)
135	3	ns	-7.14	42.86	549.00	57.59	0.95 ± 0.03	1.00 ± 0.00	0.29	***	(-1.00, 0.29)

4.7.2 Testing evolutionary histories of the parents

Coalescence analysis using Approximate Bayesian Computations. Although the expectation for the Spanish hybrid zone between *H. non-scripta* and *H. hispanica* was that they came into secondary contact after allopatric speciation, we observed large amount of shared polymorphisms in our data (Table 4.9). This genetic diversity led us to suggest alternative hypotheses. Both species might have exchanged migrants ever since their split, or the shared polymorphisms represent ancestral polymorphisms maintained due to incomplete lineage sorting because both taxa diverged relatively recently. To test the evolutionary origin of shared polymorphisms and amount of gene flow between the two populations of parental samples four likely models of evolutionary history were suggested and addressed by coalescence simulations and approximate Bayesian computations (ABC) to estimate the key parameters. A folded joint site frequency spectrum (jsfs) between pure individuals of either parental population (*H. non-scripta*: 95 individuals, *H. hispanica*: 53 individuals) was computed to generate different descriptive statistics. The genotype data were transformed so that '0' allele represented the minor allele with a frequency below 0.5 across the subset of these samples. The data subset of biSNPs for pure parental individuals (section 4.3.3) was polymorphic at only 1428 sites of which 21.3 % were largely shared. In contrast, *H. non-scripta* contained 27.4 % and *H. hispanica* 30.6 % private alleles.

Four different evolutionary histories were considered: a constant and low migration rate, m, symmetric from both parental populations since their split time T (**SyM**); asymmetric migration where m_{NStoH} corresponds to migration from the *H. non-scripta* population to the *H. hispanica* population and m_{HtoNS} to the reverse migration scheme since their split time T (**AsyM**); both populations split T_S generations ago but remained in allopatry until some constant and symmetric migration, m, established T_C generations ago (**SecM**); and lastly the ancestral population never split and unrestricted gene flow following one single panmictic population led to the current observation of shared polymorphisms (**OneSp**).

Simulations of sequence data that evolve according to a coalescent model as outlined above were carried out using *scrm()* in R (Staab et al., 2015). DNA sequences of two alleles per locus were simulated for both parental populations with their respective sample size of the real data (*H. non-scripta*: 95 individuals, *H. hispanica*: 53 individuals). The simulation replicated the marker design by sampling at least one fixed allele per gene. Further, one amplicon per gene was randomly chosen to obtain 215 loci and independence of the markers was assumed. Two populations were simulated that had diverged time T ago. Migration is implemented as the fraction of a population that is replaced with migrants from the other population each generation looking forward in time. The mutation rate (under assumption of an infinite site model) is given by θ. The same effective population size, N_e, was assumed for both parental populations and the ancestral population. All parameters were in scale to $4 \cdot N_e$. A uniform prior for all three parameters was suggested: θ (0,1), m (0,10), and T (0,5). Simulations were repeated for each model 100,000 times. For each simulation, the jsfs was re-estimated however, only its most informative entries were used, hence any statistic that almost always generated no variance was excluded. Accordingly, for the simulations only eight statistics from the jsfs were kept that generated variance across all simulations (Table 4.9).

Table 4.9 – Folded joint site frequency spectrum for 'pure' individuals of both parental species from the observed data. In bold are highlighted the statistics that were kept for the ABC estimations.

ns v hi >	$p = 0$	$p < 0.05$	$p < 0.95$	$p < 1$	$p = 1$
$p = 0$	x	**207**	**184**	0	0
$p < 0.05$	**259**	**67**	**105**	0	0
$p < 0.95$	**176**	**121**	**298**	0	0
$p < 1$	1	2	4	0	0
$p = 1$	1	1	2	0	x

Those were our statistics to predict the parameters θ, m, and T, using the ABC package in R (Csilléry et al., 2012). For each model the parameter estimates were tested by two ABC methods, the rejection algorithm and the local linear regression technique (Beaumont et al., 2002) for four different acceptance tolerance thresholds (5e-04, 0.001, 0.005, 0.01). We performed 1000 replicates of leave-one-out cross-validations to obtain a prediction error of the euclidean distance between the estimators of the parameters and their true value (Table 4.10). Model selection was performed for 1000 random replicates of leave-one-out cross-validations. The mean posterior probability is reported for each model at a given acceptance threshold, and the probability of each model given the observed jsfs (Table 4.11).

Lastly, we also tried to include hybrid samples in a model of three populations, but unfortunately estimating more parameters becomes too difficult and we did not reach enough power to make predictions here.

ABC model selection. Both the 'rejection' and 'linear correction' method (Beaumont et al., 2002) showed qualitatively the same result (not presented), but the linear correction method led to more accurate estimates, which are therefore reported in here. Though, the linear correction method was exploitable only for the most stringent threshold that keeps only the 50 best out of a hundred thousand simulations.

The cross-validation of the parameter estimates showed relatively small prediction errors for θ (< 0.006) and m (< 0.1), but less good estimates of T (< 0.65) (Table 4.10). Noticeably, for both models with four parameters to be estimated (AsyM and SecM), the prediction error increased with more stringent acceptance thresholds (Table 4.10). The prediction error compared the mean posterior to the true value while ignoring the shape of the posterior itself. Hence, ABC was performed for all different tolerance thresholds.

Based on cross-validation of posterior model probabilities, the most likely model among the four proposed evolutionary histories was the model with constant symmetric migration since the split between both parental populations (Data to SyM, Table 4.11). This was indicated by the highest proportions of 1000 replicates from oberved data (i.e. 97.8 %) that best fit to the SyM model. Despite that, the power to distinguish between the different models themselves was very low, e.g. maintaining only 0.05 % of all simulations from model SyM, only 46.1 % of the replicates were detected as model SyM in contrast to the other two models (SyM, Table 4.11).

Table 4.10 – Mean prediction error from cross validation of 1000 randomly drawn posterior parameter estimates for all four models. The smaller the prediction errors, the closer the estimates are to the true value.

(a) SyM

Threshold	θ	m	T
$5e-04$	0.005338524	0.062785543	0.431851775
0.001	0.004891097	0.053289617	0.395122825
0.005	0.004670313	0.048725401	0.369894225
0.01	0.004720287	0.048317868	0.367569271

(b) AsyM

Threshold	θ	m_{NStoH}	m_{HtoNS}	T
$5e-04$	0.005397950	0.097396962	0.101782729	0.650207589
0.001	0.004984305	0.088781288	0.094099577	0.585499519
0.005	0.004722706	0.084476819	0.086705211	0.544511605
0.01	0.004715957	0.085184305	0.086808435	0.544098994

(c) SecM

Threshold	θ	m	T_S	T_C
$5e-04$	0.005586149	0.076121971	0.487835773	0.138894270
0.001	0.005560154	0.074508102	0.477187755	0.136982685
0.005	0.006028528	0.082786325	0.506200311	0.138924392
0.01	0.006354110	0.084582148	0.515472887	0.143049159

(d) OneSp

Threshold	θ
$5e-04$	0.005866036
0.001	0.005762230
0.005	0.006114518
0.01	0.006264786

Table 4.11 – Mean model posterior probabilities over 1000 replicates depicted for the following thresholds of maintained fraction of simulations after linear correction on the parameter estimates: 0.5 %, 0.1 %, and 0.05 %. The last row reports the probability of each model given the jsfs statistics in the observed data. The OneSp model was actually never chosen and is therefore not shown.

	SyM			**AsyM**			**SecM**		
threshold	0.5 %	0.1 %	0.05 %	0.5 %	0.1 %	0.05 %	0.5 %	0.1 %	0.05 %
SecM	0.3105	0.3218	0.3235	0.1847	0.1735	0.1553	0.5048	0.5046	0.5212
AsyM	0.1849	0.1789	0.1808	0.6572	0.6862	0.6794	0.158	0.135	0.1399
SyM	0.4326	0.4601	0.461	0.2284	0.218	0.2155	0.339	0.3219	0.3235
Data	0.585	0.8018	0.9779	0.0547	0.1696	0.0053	0.3603	0.0286	0.0168

Table 4.12 – Parameter estimates based on the symmetric migration model since species split (**SyM** model) with 0.05 % threshold with the linear correction method.

	θ	m	T
2.5%	0.5439	1.3628	1.0126
mean	0.5504	1.4160	1.2311
97.5%	0.5568	1.4554	1.4327

Estimate of split time, effective population size, and migrants. As mentioned in section 4.3.1, the average size of an amplicon was 131.5 bp. Using the parameter estimates of the best fit model (SyM) obtained after local linear correction at a threshold of 0.05 % the mean θ (0.5504, Table 4.12) calculated per base pair was $4.19 \cdot 10^{-3}$ (2.5 %: $4.14 \cdot 10^{-3}$; 97.5 %: $4.23 \cdot 10^{-3}$). This is coherent with the estimate of π from the transcriptome assembly ($\pi_{ns} = 5.6 \cdot 10^{-3}$, chapter 3).

Using those estimates and assuming a mutation rate per site of $\mu = 10^{-8}$, we can infer the effective population size, N_e, to be around $1.05 \cdot 10^6$ ($1.03 \cdot 10^6 - 1.06 \cdot 10^6$). In turn, this enumerated a migration rate per generation of approximatively $0.338 \cdot 10^{-6}$ ($0.326 \cdot 10^{-6} - 0.348 \cdot 10^{-6}$).

Lastly, the split between both bluebells depends on the assumption of a generation time for bluebells. They start flowering at a critical dry bulb mass of 0.2 - 0.4 g (Blackman and Rutter, 1954), and Woodhead (1904) reported that they flower first in their fifth year. Because each season the old bulb is replaced by a new one, but benefiting from the old's resources an 'individual' plant can live up to many years (Al-Modayan, 1993). Ferriday (1987) estimated the average life-span to be 16 years. Consequently, the age distance between parents and their offspring can range from five to more than 16 years. Assuming a minimum generation time of five years, the split time enumerates as 1.15 mya (1.12 mya - 1.18 mya) or assuming 16 years the split time is 3.69 mya (3.6 mya – 3.77 mya). The latter estimate might be more appropriate given that bluebells also reproduce clonally.

Low migration rate and recent split time suggests that the large amount of shared polymorphisms are most likely due to inherited ancestral variation and not a secondary contact. Both split time estimates are within the range of the Pleistocene epoch and would support the hypothesis of parapatric speciation due to climatic oscillations. The alternative hypothesis that *H. non-scripta* and *H. hispanica* would represent one panmictic population was not supported at all. Overall, the ABC results are interesting but they might also be somewhat unrealistic. Previous analyses of local population structure suggest differentiation between collecting sites, which are ignored by the applied models. Simulating more populations however was (so far) not possible.

Chapter 5

Concluding remarks

5.1 Aims of these studies

One of the motivations for this PhD was the conservation interest in the British bluebell, *Hyacinthoides non-scripta*, in the British Isles, where it is put at risk by invasion and introgressive hybridisation with its conspecific *H. hispanica* and bluebell garden varieties. In addition, a natural hybrid zone between *H. non-scripta* and *H. hispanica* was described in north-western Iberian Peninsula that provided the opportunity to explore the extent of inter-specific hybridisation, the direction of gene flow, and potential reproductive barriers in a natural laboratory. Such information will be valuable for comparison to the British situation, which is under anthropogenic influence. Crossing experiments between living samples from the hybrid zone were successfully used to obtain breeding system information for each taxon and to explore the hybrid fitness and chance of hybrid formation. In addition, genome size estimates and morphological characterisation presented valuable hybrid features. We identified a lack of genome-wide markers to study hybridisation in the two species with large genomes. Accordingly, in this thesis the genomes of both species were unlocked by designing amplicon sequences from exons that were re-sequenced using multiplexing PCR technology for hundreds of individuals. Further specifications for selecting the amplicon sequences were applied to provide species specific, hence diagnostic, markers. Population genetic analyses and cline analysis on the re-sequencing data were lastly performed to explore the origin of the hybrid zone, the amount and direction of introgression and reproductive barriers.

5.2 Homoploid hybrid zone with phenotypic intermediate hybrids and their breeding system

The analysis of genetic markers of hybrid zones need to be interpreted in the light of ecological and biological properties of the interacting species. Especially for non-model organisms such information has not been systematically obtained and complementing studies need to be conducted. Many resources are available for *H. non-scripta*, as it is an iconic species of public interest in the British Isles. In contrast, its Iberian close relative, *H. hispanica* has received little attention. For instance, self-incompatibility of *H. non-scripta* has been studied at least three times throughout the last century (Corbet,

1998; Knight, 1964; Wilson, 1959b), while nothing is known about the breeding system of *H. hispanica*. Hybridisation between both species has been noted in several regions of northern Europe (e.g. Geerinck, 1996; Ietswaart et al., 1983; Page, 1987). The frequency at which hybrids are formed between both taxa and the hybrid fitness, however, has not been studied so far. This chapter intended to close some information gaps that would impact the interpretation of interactions between both species.

Intensive field sampling of 39 different collecting sites was performed in the Galician-Duero Mountains and surrounding areas spanning 120 km north to south and about 90 km from west to east. Occurrences of bluebells were found between 500 – 1200 m altitude in mostly deciduous *Quercus* forests, which are the dominant plant society of northern Iberia since the Younger Dryas (11.7 ky B.P. Sobrino et al., 2007). The possibility of a secondary contact between *H. non-scripta* and *H. hispanica* within northwestern Iberia was discussed as a consequence of the last glacial maximum.

Genome size estimates were significantly different between both species. *H. non-scripta* from northern Spain presented a smaller genome size with 2C = 47.44 pg, and *H. hispanica* from its southern range in Portugal was larger 2C = 49.63 pg. Contrarily to reports of triploid individuals of either taxon from the UK (see supplement of Grundmann et al., 2010), there was no evidence for polyploid plants amongst the new Iberian samples. Further, the hybrids showed intermediate genome sizes between the two parental taxa and therefore a homoploid hybrid zone was concluded.

Morphological documentations from the field and scoring of flowers *ex situ* evidenced a range of intermediate phenotypes of the hybrids. Overall, there was a gradual transition of phenotypes along latitude from one parental taxon to the other. However, the hybrids resembled the parental species they were more closely to in the field, probably due to the geological separation of the hybrids by the mountain range from the Sierra del Teleno to the Montes Aquilianos. Consequently, the hybrids were grouped into one northern and one southern hybrid population.

Lastly, crossing experiments were performed that showed strong self-incompatibility for the hybrids and *H. hispanica*. In *H. non-scripta* more outliers of flowers that set seeds were discovered, which could present a slightly less strong self-incompatibly mechanism. Additionally, seeds obtained from selfing crosses germinated successfully without strong differences compared to outcrossed seeds, despite difficulties with the optimal germination protocol. The inter-specific crosses between *H. non-scripta* and *H. hispanica* that produced F_1 seedlings resulted in slightly higher proportions of seeds per fruit than outcrosses. This effect was only significant for inter-specific crosses, in which *H. hispanica* was the ovule donor. The hybrid crosses revealed that they can produce as many seeds as their parents, and their seeds also germinated successfully. Additionally, the proportion of produced seeds per flower were slightly increased in the outbreeding crosses, with significant differences only found in the hybrid South population. These results suggests overall a low reproductive isolation between *H. non-scripta* and *H. hispanica* within some limitations, as the crossing experiments ignore pre-zygotic reproductive barriers and do not assess the long-term survival and fitness of the F_1 hybrids (i.e. post-zygotic reproductive barrier). Neverthelees, the hybrids showed high fecundity with outbreeding tendencies.

Accordingly, expectations for the natural bluebell hybrid zone were large amount of

introgression, few loci that present reproductive barriers and a possible heterozygote fitness advantage in hybrids.

5.3 Development of a marker set suitable for hybridisation studies of non-model species with large genomes

For model or crop organisms (for example arabidopsis, rice, banana) many genomic resources are available, including assembled genomes, annotated genes, transcriptomes and population genetics markers. But for the majority of species few or no such resources exist. When the genome is very large (1C > 14 Gb), genomic resources remain very costly or challenging to obtain, despite decreasing costs of next-generation sequencing. The purpose of this chapter was to establish a bioinformatics pipeline that will enable population genetic studies with large numbers (hundreds) of novel genomic markers at modest cost for species with very large genomes. To achieve these ends, *de novo* assembled RNAseq data was used to extract shared genes between three bluebell species, in genus *Hyacinthoides* Heist. ex Fabr. Using the assembled transcriptomes, a targeted exon re-sequencing strategy was developed, through which 300 primer pairs were identified that contain potentially informative single nucleotide polymorphisms in 221 nuclear genes, ten chloroplast and five mitochondrial genes. Using Fluidigm technology, these primer pairs were then used to amplify genomic DNA from individuals of the British bluebell, *H. non-scripta* from the UK, France and Spain and *H. hispanica* from the Iberian Peninsula. Both species have large genome sizes (1C = 23.2 and 24.3 Gb, respectively). The approach worked effectively and produced (few diagnostic) markers to differentiate both species that are reproducible at low costs. The markers therefore enable future conservation studies addressing the degree of introgression of continental bluebell species into British *H. non-scripta*, and fundamental studies tracing genetic markers across a bluebell hybrid zone in northern Spain. More generally, the pipeline is widely applicable to any population genetics study that is targeting species with large genomes.

5.4 Reproductive isolation despite long-term gene flow across a natural hybrid zone

Hybrid zone studies are well suited to explore the origin of speciation and its drivers. Across narrow regions where two distinguishable populations interbreed and form hybrids, patterns in molecular markers can predict the past population dynamics, present asymmetries in introgression, and reveal reproductive barriers. In this chapter the genetic transition between *H. non-scripta* and *H. hispanica* has been studied along a natural hybrid zone in the north-west of the Iberian Peninsula. The molecular patterns provided multiple hypotheses of forces driving divergence between the two species in the Galician-Duero Mountains. First of all, the hybrid zone was suggested to represent primary intergradation from a parapatric speciation. The species split was estimated for 1.15 – 3.69 mya with very small (but constant) amount of migrants ever since the split. The climatic Quaternary oscillations and the complex geographic area of the Cantabrian Mountains and the inner Galician-Duero Mountains along a transition of two biogeographical regions, the

Eurosiberian and Mediterranean, have potentially played their part in divergent evolution despite gene flow and maintaining hybrid populations. A relatively large amount of shared polymorphisms has been recovered in the re-sequencing data of 307 samples, which are probably remnants of ancestral common heritage. The possibility of contemporary hybridisation is certainly given on the basis of assumed long-distance pollen dispersal, the crossing experiments, and backcrossed individuals presented in the southern hybrid population. However, early generations of hybrids (F_1, F_2) were concluded to be rare because no sympatric parental populations were discovered in the collected sites. Gene flow between individual bluebells and collecting sites was assumed to be mainly mediated through pollen dispersal because their seed dispersal is very limited. Consequently, gene flow across the whole of the hybrid zone should require stepwise sequence of introgression, i.e. from *H. non-scripta* to hybrid North, to hybrid South, to *H. hispanica* (and/or in reverse sequence), and rare long-distance dispersal of pollen. The patterns of differential introgression between nuclear and plastid markers were interpreted as a potential southwards movement of *H. non-scripta* into *H. hispanica*. It was reasoned for by the high organelle haplotype frequency of *H. hispanica* amongst hybrid individuals even far north and consequently *H. hispanica* must have been distributed further north at some former time. The movement of the hybrid zone could be driven by asymmetrical hybridisation (evidenced by the F_1 hybrid advantage of seed crosses for *H. hispanica* mothers, and stronger isolation of northern hybrids to *H. non-scripta*), differences in flowering time (*H. hispanica* flowered later and could be swamped by pollen of hybrid North but rarely of *H. non-scripta* directly), and climate change. There were also indications of reproductive isolation by steep clines of nuclear SNPs, which were concordant with the organelle cline suggesting cyto-nuclear incompatibilities. The evolution of such cyto-nuclear incompatibilities were argued to drive divergence despite gene flow by adaptive advantage of organelle haplotype to their local environment, and nuclear genotypes would slowly follow the pattern. Several of these hypotheses need to be complimented by additional analyses. Nonetheless, the results were surprising in that initially the hybrid zone was considered a consequence of secondary contact.

5.5 Outlook – genomic pollution by introgressive hybridisation in the UK

The same marker set has been re-sequenced for samples from the UK by collaborators of the Royal Botanic Garden Edinburgh with the aims to identify hybrid individuals, and to trace the advance of hybridisation between the alien taxa and the native *H. non-scripta*. The nature of the genetic markers pose some challenges as they are rather slowly evolving exonic markers and they have revealed rather old hybridisation and shared ancestral polymorphisms in Iberian samples. The first introduction of *H. hispanica* to the British Isles, on the other side, was first reported in 1683 (Pilgrim and Hutchinson, 2004). Therefore, hybridisation between alien taxa and the native *H. non-scripta* can be expected to represent a much more recent pattern. The diagnostic markers will reliably identify early generations of hybrids. However, local substructure between collecting sites might be missed due to a lack of singletons and other private alleles of the slow marker system. To

clarify, a rapid expansion of *H. non-scripta* from some southern refugia (northern side of the Cantabrian Mountains or further North) is commonly accompanied by a strong bottleneck, which reduces genetic diversity through the survival of only a few genetic lineages (Hewitt, 1996). Similarly, human introduction of a few individuals of *H. hispanica* or garden variants lead to a so-called founder effect (i.e. due to a small number of immigrants only a small proportion of the genetic diversity of the source population is present, Lee, 2002), which represents an even stronger bottleneck. Accordingly, the reduced genetic diversity in the parental taxa of the UK hybridisation might cause an underestimation of local (meta-)population substructure. For example, the clustering results of *H. non-scripta* in chapter 3 showed only a low genetic differentiation between Spanish samples from those of French or British origin. The benefit of the slowly evolving marker system is that they promise to reliably amplify and hinder null alleles, as demonstrated in the *proof of concept* of re-sequencing samples of both parental species across their distribution range (chapter 3).

Given the potential parapatric origin of hybridisation in Iberia, the genomic patterns for secondary contact in the British Isles might be different: 1) The genetic nature of the 'Spanish bluebell' has not been clarified yet in relation to the gene pool for *H. hispanica*. However, the plastid phylogeny by Grundmann et al. (2010) suggested a southern Portuguese origin. Our genetic marker system successfully differentiated southern samples (BB-188, BB-262), and additionally captured a third organelle haplotype, which could be related to a Spanish cultivar sample (chapter 3). 2) The frequency of *H. non-scripta* samples in native bluebell forests of Britain should outnumber the invading samples and (demographic) swamping of the invading taxa can be assumed by more abundant *H. non-scripta* pollen (Todesco et al., 2016; Wolf et al., 2001). However, the low reproductive barrier can lead to extensive hybridisation, which in combination with stronger hybrid fitness can lead to genetic swamping of both taxa and lead to their replacement by hybrids (Todesco et al., 2016). Such adaptive introgressive hybridisation could be beneficial to *H. non-scripta* by facilitating adaptation to the rapidly warming climate change through increased genetic variance (Brennan et al., 2015; Lee, 2002) at the cost of some beloved phenotypic features (scent, drooping flowers). Therefore, the hybrid fitness and reproductive isolation will be important in this system. For instance, heterotic hybrids and *H. hispanica* plants have been reported in the UK (Wilson, 1956, 1958), as well as, triploid *H. non-scripta* and *H. hispanica* that would pose a strong reproductive barrier, if they present unbalanced, nonfunctional gametes resulting in triploid sterility (Ramsey and Schemske, 1998). But it is also possible that they produce balanced gametes and present higher fitness due to heterosis (Chapman and Abbott, 2010). The feasibility of the marker system for polyploid samples would still need to be explored and the variant discovery method would need to be adapted accordingly. The potential triploid samples from chapter 3 were not attempted to analyse further, but the presence of multiple (> 2) alleles was restricted to a few amplicon regions. The PCR amplicon re-sequencing provided a high read depth but the read depth also varied strongly between loci or amplicons, and samples. A genotyping method for polyploid samples that is based on read depth such as developed by Zohren et al. (2016) assumes unbiased and independent read sampling. 3) The cyto-nuclear incompatibilities in the hybrid zone might not be present in Britain's bluebell taxa due

specific adaptive forces of environmental/climate change of northern Iberia. However, a transition of flowering time along latitude might pose a reproductive barrier (pers. comm Fred Rumsey).

Appendix A

Supplementary tables

Table A.1 – List of individuals for which genome size measurements were obtained in chapter 2, and which individuals were sequenced in chapter 3 and 4. 'CPG' in the sample's name indicates it origin from living material at the Chelsea Physics Garden, London. For each sample the organellar haplotype (**CP**) is given with 0 – *H. non-scripta*, 1 – *H. hispanica*, and 3 – exclusive to site BB-262. In addition, the **accession** number from the NHM (if available) was listed along with whether DNA was obtained from **dried** silica gel or **fresh** leaf material, and the obtained genome size (**Gs**). For the genome size the asteriks (*) indicates when an estimate's CV is below 5 %. For chapter 3 it is highlighted if the sample was used only for the **org**anellar genotyping, or for the organellar and nuclear (**both**) genotyping, or had its transcriptome sequenced (**mRNA**).

Species	Individual	Pop	CP	Accession	Tissue	Gs	Chapter 3	Chapter 4
H. non-scripta	126-2	126	0	BM000865085	dried		org	–
H. non-scripta	126-9	126	0	BM000865085	dried		org	–
H. non-scripta	135-10	135	0	BM000865094	dried		both	x
H. non-scripta	135-11	135	0	BM000865094	dried		both	x
H. non-scripta	135-7	135	0	BM000865094	dried		both	x
H. hispanica	188-B-31-CPG	188	1	BM000865394	fresh	49.63*	both	–
H. hispanica	262-B-01-CPG	262	3	BM000865511	fresh		org	–
H. hispanica	262-B-33-CPG	262	3	BM000865511	fresh		both	–
H. non-scripta	346-05-CPG	346	0	BM000864254	fresh		both	–
H. non-scripta	347-02-CPG	347	0	BM000864254	fresh		both	–
H. non-scripta	347-04-CPG	347	0	BM000864254	fresh		both	–
H. non-scripta	347-05-CPG	347	0	BM000864254	fresh		both	–
H. hispanica	353-02-CPG	353	1	BM000864261	fresh		org	–
H. hispanica	361-1	361	1	BM000864269	dried		both	x
H. hispanica	361-2	361	1	BM000864269	dried		both	x
H. hispanica	361-3	361	1	BM000864269	dried		both	x
H. hispanica	361-4	361	1	BM000864269	dried		both	x
H. hispanica	361-5	361	1	BM000864269	dried		both	x
H. non-scripta	391-1	391	0	BM000864300	dried		both	–
H. non-scripta	391-2	391	0	BM000864300	dried		both	–
H. non-scripta	391-3	391	0	BM000864300	dried		both	–
H. non-scripta	391-4	391	0	BM000864300	dried		both	–
H. non-scripta	391-5	391	0	BM000864300	dried		both	–
H. non-scripta	392-04-CPG	392	NA	BM000864301	fresh	48.65*	–	–
H. non-scripta	392-05-CPG	392	0	BM000864301	fresh		both	–
H. non-scripta	393-1	393	0	BM000864302	dried		both	x
H. non-scripta	393-2	393	0	BM000864302	dried		both	x

Table A.1 continued

Species	Individual	Pop	CP	Accession	Tissue	Gs	Chapter 3	Chapter 4
H. non-scripta	393-3	393	0	BM000864302	dried		both	x
H. non-scripta	393-3-CPG	393	0	BM000864302	fresh		both	–
H. non-scripta	393-4	393	0	BM000864302	dried		both	x
H. non-scripta	393-5	393	0	BM000864302	dried		both	x
H. non-scripta	395-2	395	0		dried		–	x
H. non-scripta	395-3	395	0		dried		–	x
H. non-scripta	395-7	395	0		dried		–	x
H. non-scripta	395-8	395	0		dried		–	x
H. non-scripta	395-A	395	0		fresh	45.96*	–	x
H. non-scripta	395-B	395	0		fresh		–	x
H. non-scripta	395-C	395	0		fresh		–	x
H. non-scripta	397-1-CPG	397	0	BM000864306	fresh	48.00*	both	x
H. non-scripta	397-2-CPG	397	0	BM000864306	fresh	48.00*	both	x
H. non-scripta	397-3	397	0	BM000864306	dried		both	x
H. non-scripta	397-4-CPG	397	0	BM000864306	fresh		both	x
H. non-scripta	397-5	397	0	BM000864306	dried		both	x
H. non-scripta	398-1	398	0	BM000864307	dried		both	x
H. non-scripta	398-2	398	0	BM000864307	dried		both	x
H. non-scripta	398-3	398	0	BM000864307	dried		both	x
H. non-scripta	398-4	398	0	BM000864307	dried		both	x
H. non-scripta	398-5	398	0	BM000864307	dried		both	x
hybrid North	400-10	400	1		dried		–	x
hybrid North	400-4	400	0		dried		–	x
hybrid North	400-7	400	1		dried		–	x
hybrid North	400-9	400	1		dried		–	x
hybrid North	400-A	400	1		fresh	45.96*	–	x
hybrid North	400-B	400	0		fresh	47.02	–	x
hybrid North	400-C	400	1		fresh	46.39	–	x
hybrid North	401-4	401	0		dried		–	x
hybrid North	401-6	401	0		dried		–	x
hybrid North	401-9	401	0		dried		–	x
hybrid North	401-A-wh	401	0		fresh	45.97*	–	x
hybrid North	401-B	401	0		fresh	46.43*	–	x
hybrid North	401-C	401	0		fresh	46.09*	–	x
hybrid North	401-D	401	0		fresh	46.34	–	x
hybrid North	402-14	402	1		dried		–	x
hybrid North	402-15	402	1		dried		–	x
hybrid North	402-9	402	1		dried		–	x
hybrid North	402-A	402	0		fresh	46.3	–	x
hybrid North	402-B	402	1		fresh	46.32*	–	x
hybrid North	402-C	402	NA		fresh	46.33	–	–
hybrid North	402-D	402	0		fresh	45.77*	–	x
hybrid North	402-E	402	0		fresh	45.90*	–	x
hybrid North	402-F	402	NA		fresh	45.73*	–	–
hybrid North	403-8	403	1		dried		–	x
hybrid North	403-A	403	0		fresh	45.73*	–	x
hybrid North	403-B	403	1		fresh	46.46*	–	x
hybrid North	403-C	403	0		fresh	46.22*	–	x
hybrid North	403-D	403	1		fresh	46.05	–	x
hybrid North	403-E	403	0		fresh	46.45*	–	x
hybrid North	403-F	403	1		fresh		–	x
H. non-scripta	404-1	404	0		dried		–	x

Table A.1 continued

Species	Individual	Pop	CP	Accession	Tissue	Gs	Chapter 3	Chapter 4
H. non-scripta	404-2	404	0		dried		–	x
H. non-scripta	404-3	404	0		dried		–	x
H. non-scripta	404-5	404	0		dried		–	x
H. non-scripta	404-6	404	0		dried		–	x
H. non-scripta	405-1-B	405	0		fresh	46.69	–	x
H. non-scripta	405-10-8	405	0		dried		–	x
H. non-scripta	405-11-C	405	0		fresh	46.38	–	x
H. non-scripta	405-5-5	405	0		dried		–	x
H. non-scripta	405-6-7	405	0		dried		–	x
H. non-scripta	405-7-A	405	0		fresh	46.81	–	x
H. non-scripta	405-9-6	405	0		dried		–	x
H. non-scripta	406-3	406	0		dried		–	x
H. non-scripta	406-3-NHM	406	0		dried		–	x
H. non-scripta	406-4-NHM	406	0		dried		–	x
H. non-scripta	406-5-NHM	406	0		dried		–	x
H. non-scripta	406-6	406	0		dried		–	x
H. non-scripta	406-8	406	0		dried		–	x
H. non-scripta	406-A-7	406	0		fresh	2n	–	x
hybrid South	407-1	407	1		dried		–	x
hybrid South	407-2	407	1		dried		–	x
hybrid South	407-3	407	1		dried		–	x
hybrid South	407-4	407	1		dried		–	x
hybrid South	407-5	407	1		dried		–	x
H. non-scripta	410-2-NHM	410	0	BM000864319	dried		both	x
H. non-scripta	410-4-NHM	410	0	BM000864319	dried		both	x
H. non-scripta	410-5-NHM	410	0	BM000864319	dried		both	x
H. non-scripta	413-2-CPG	413	0	BM000864322	fresh		both	–
H. non-scripta	415-4-CPG	415	0	BM000864324	fresh		both	–
H. hispanica	460-1	460	1	BM000864368	dried		both	x
H. hispanica	460-2	460	1	BM000864368	dried		both	x
H. hispanica	460-4	460	1	BM000864368	dried		both	x
H. hispanica	460-5	460	1	BM000864368	dried		both	x
H. hispanica	460-6	460	1	BM000864368	dried		both	x
H. hispanica	460-7	460	1	BM000864368	dried		both	x
H. hispanica	461-1	461	1	BM000864369	dried		both	x
H. hispanica	461-2	461	1	BM000864369	dried		both	x
H. hispanica	461-3	461	1	BM000864369	dried		both	x
H. hispanica	461-4	461	1	BM000864369	dried		both	x
H. hispanica	461-5	461	1	BM000864369	dried		both	x
H. hispanica	461-6	461	1	BM000864369	dried		both	x
H. hispanica	461-8	461	1	BM000864369	dried		both	x
H. hispanica	462-11	462	1		dried		–	x
H. hispanica	462-4	462	1		dried		–	x
H. hispanica	462-6	462	1		dried		–	x
H. hispanica	462-9	462	1		dried		–	x
H. hispanica	462-A	462	1		fresh	48.76	–	x
H. hispanica	462-B	462	1		fresh	48.38	–	x
H. hispanica	462-C	462	1		fresh	47.8	–	x
H. non-scripta	472-4	472	0		dried		–	x
H. non-scripta	472-5	472	0		dried		–	x
H. non-scripta	472-6	472	0		dried		–	x
H. non-scripta	472-7	472	0		dried		–	x

Table A.1 continued

Species	Individual	Pop	CP	Accession	Tissue	Gs	Chapter 3	Chapter 4
H. non-scripta	472-A	472	0		fresh	48.36*	–	x
H. non-scripta	472-B	472	0		fresh	48.39*	–	x
H. non-scripta	472-C	472	0		fresh	47.78*	–	x
H. non-scripta	474-2	474	0		dried		–	x
H. non-scripta	474-3	474	0		dried		–	x
H. non-scripta	474-4	474	0		dried		–	x
H. non-scripta	474-7	474	0		dried		–	x
H. non-scripta	474-8	474	0		dried		–	x
H. non-scripta	474-9	474	0		dried		–	x
H. non-scripta	474-A	474	0		fresh	47.96*	–	x
H. non-scripta	475-1-A	475	0		fresh	48.16	–	x
H. non-scripta	475-2-8	475	0		dried		–	x
H. non-scripta	475-2-B	475	0		fresh	49.33*	–	x
H. non-scripta	475-4-7	475	0		dried		–	x
H. non-scripta	475-5-11	475	0		dried		–	x
H. non-scripta	475-5-9	475	0		dried		–	x
H. non-scripta	475-5-C	475	0		fresh	47.74*	–	x
H. non-scripta	476-A	477	0		fresh	48.37	–	x
H. non-scripta	477-A	477	0		fresh	48.56	–	x
H. non-scripta	477-B	477	0		fresh		–	x
hybrid South	479-10	479	1		dried		–	x
hybrid South	479-4	479	1		dried		–	x
hybrid South	479-5	479	0		dried		–	x
hybrid South	479-6	479	0		dried		–	x
hybrid South	479-7	479	1		dried		–	x
hybrid South	479-8	479	0		dried		–	x
hybrid South	479-A	479	0		fresh	48.3	–	x
hybrid South	480-4	480	1		dried		–	x
hybrid South	480-6	480	1		dried		–	x
hybrid South	480-7	480	0		dried		–	x
hybrid South	480-8	480	1		dried		–	x
hybrid South	480-A	480	1		fresh	48.26	–	x
hybrid South	480-B	480	1		fresh	48.61	–	x
hybrid South	480-C	480	1		fresh	48.67	–	x
hybrid South	481-5	481	0		dried		–	x
hybrid South	481-6	481	0		dried		–	x
hybrid South	481-7	481	1		dried		–	x
hybrid South	481-A	481	1		fresh	47.90*	–	x
hybrid South	481-B	481	1		fresh	48.71	–	x
hybrid South	481-C	481	1		fresh	48.33*	–	x
hybrid South	481-D	481	1		fresh		–	x
hybrid South	482-11	482	1		dried		–	x
hybrid South	482-6	482	1		dried		–	x
hybrid South	482-8	482	1		dried		–	x
hybrid South	482-9	482	1		dried		–	x
hybrid South	482-A	482	1		fresh	49.31	–	x
hybrid South	482-B	482	1		fresh	48.92	–	x
hybrid South	482-C	482	1		fresh	49.43	–	x
H. hispanica	483-8	483	1		dried		–	x
H. hispanica	483-9	483	1		dried		–	x
H. hispanica	483-A	483	1		fresh	48.02*	–	x
H. hispanica	483-B	483	1		fresh	48.07	–	x

Table A.1 continued

Species	Individual	Pop	CP	Accession	Tissue	Gs	Chapter 3	Chapter 4
H. hispanica	483-C	483	1		fresh	48.78	–	x
H. hispanica	483-D	483	1		fresh	48.30*	–	x
H. hispanica	483-E	483	1		fresh	48.57	–	x
H. hispanica	484-10-9	484	1		dried		–	x
H. hispanica	484-2-5	484	1		dried		–	x
H. hispanica	484-4-7	484	1		dried		–	x
H. hispanica	484-8-12	484	1		dried		–	x
H. hispanica	484-B	484	1		fresh	48.19	–	x
H. hispanica	484-C	484	1		fresh	48.36	–	x
H. hispanica	484-D	484	1		fresh	50.2	–	x
H. hispanica	484b-A	484	NA		fresh		–	–
hybrid North	486-1-6	486	0		dried		–	x
hybrid North	486-1-7	486	1		dried		–	x
hybrid North	486-3-4	486	1		dried		–	x
hybrid North	486-3-9	486	1		dried		–	x
hybrid North	486-4-10	486	0		dried		–	x
hybrid North	486-4-3	486	0		dried		–	x
hybrid North	486-A	486	1		fresh	48.54	–	x
hybrid North	487-4	487	0		dried		–	x
hybrid North	487-5	487	0		dried		–	x
hybrid North	487-6	487	0		dried		–	x
hybrid North	487-7	487	0		dried		–	x
hybrid North	487-8	487	1		dried		–	x
hybrid North	487-9	487	1		dried		–	x
hybrid North	487-A	487	0		fresh	50.32	–	x
hybrid North	488-3	488	1		dried		–	x
hybrid North	488-4	488	1		dried		–	x
hybrid North	488-A	488	1		fresh		–	x
hybrid North	488-B-6	488	0		fresh		–	x
hybrid North	488-C-5	488	1		fresh		–	x
H. non-scripta	489-10	489	0		dried		–	x
H. non-scripta	489-3	489	0		dried		–	x
H. non-scripta	489-4	489	0		dried		–	x
H. non-scripta	489-6	489	0		dried		–	x
H. non-scripta	489-A	489	0		fresh	48.54*	–	x
H. non-scripta	489-B	489	0		fresh		–	x
H. non-scripta	489-C	489	0		fresh	48.38	–	x
H. hispanica	490-10	490	1		dried		both	x
H. hispanica	490-4	490	1		dried		both	x
H. hispanica	490-5	490	1		dried		both	x
H. hispanica	490-6	490	1		dried		both	x
H. hispanica	490-7	490	1		dried		both	x
H. hispanica	490-A	490	1		fresh	48.47*	both	x
H. hispanica	490-C	490	NA		fresh	48.11*	–	–
H. hispanica	490-C	490	1		fresh	48.27	both	x
H. hispanica	491-11	491	1		dried		–	x
H. hispanica	491-5	491	1		dried		–	x
H. hispanica	491-6	491	1		dried		–	x
H. hispanica	491-8	491	1		dried		–	x
H. hispanica	491-A	491	1		fresh	48.29*	–	x
H. hispanica	491-B	491	1		fresh	48.29	–	x
H. hispanica	491-C	491	1		fresh	48.62*	–	x

Table A.1 continued

Species	Individual	Pop	CP	Accession	Tissue	Gs	Chapter 3	Chapter 4
hybrid South	492-6	492	0		dried		–	x
hybrid South	492-7	492	1		dried		–	x
hybrid South	492-A	492	0		fresh	48.73	–	x
hybrid South	492-B	492	NA		fresh	48.83*	–	–
hybrid South	492-C	492	0		fresh	48.31*	–	x
hybrid South	492-D	492	1		fresh	47.97*	–	x
hybrid South	492-E	492	1		fresh	48.51”	–	x
hybrid South	493-10	493	1		dried		–	x
hybrid South	493-2	493	1		dried		–	x
hybrid South	493-6	493	1		dried		–	x
hybrid South	493-9	493	1		dried		–	x
hybrid South	493-A	493	1		fresh	48.3	–	x
hybrid South	493-B	493	1		fresh	48.01	–	x
hybrid South	493-C	493	1		fresh	49.05	–	x
H. hispanica	494-11	494	1		dried		–	x
H. hispanica	494-13	494	1		dried		–	x
H. hispanica	494-4	494	1		dried		–	x
H. hispanica	494-9	494	1		dried		–	x
H. hispanica	494-A	494	1		fresh		–	x
H. hispanica	494-B	494	1		fresh		–	x
H. hispanica	494-C	494	1		fresh		–	x
hybrid North	495-4	495	0		dried		–	x
hybrid North	495-5	495	1		dried		–	x
hybrid North	495-8	495	1		dried		–	x
hybrid North	495-9	495	1		dried		–	x
hybrid North	495-A	495	0		fresh	48.48	–	x
hybrid North	495-B	495	0		fresh	48.6	–	x
hybrid North	495-C	495	0		fresh	48.59	–	x
hybrid North	496-11	496	1		dried		–	x
hybrid North	496-5-1	496	1		dried		–	x
hybrid North	496-5-2	496	1		dried		–	x
hybrid North	496-7	496	0		dried		–	x
hybrid North	496-9	496	0		dried		–	x
hybrid North	496-A	496	1		fresh	48.52*	–	x
hybrid North	496-B	496	0		fresh	49.5	–	x
hybrid South	497-11	497	1		dried		–	x
hybrid South	497-12	497	0		dried		–	x
hybrid South	497-5	497	1		dried		–	x
hybrid South	497-6	497	1		dried		–	x
hybrid South	497-8	497	1		dried		–	x
hybrid South	497-9	497	1		dried		–	x
hybrid South	497-A	497	1		fresh	47.53*	–	x
hybrid South	497-B	497	NA		fresh	47.78*	–	–
hybrid South	497-C	497	NA		fresh	48.66*	–	–
hybrid South	497-D	497	NA		fresh	48.20*	–	–
hybrid South	498-11	498	1		dried		–	x
hybrid South	498-4	498	1		dried		–	x
hybrid South	498-5	498	1		dried		–	x
hybrid South	498-6	498	1		dried		–	x
hybrid South	498-A	498	1		fresh	48.91	–	x
hybrid South	498-B	498	1		fresh	46.69*	–	x
hybrid South	498-C	498	1		fresh	45.76	–	x

Table A.1 continued

Species	Individual	Pop	CP	Accession	Tissue	Gs	Chapter 3	Chapter 4
H. hispanica	499-4	499	1		dried		–	x
H. hispanica	499-5	499	1		dried		–	x
H. hispanica	499-6	499	1		dried		–	x
H. hispanica	499-7	499	1		dried		–	x
H. hispanica	499-A	499	1		fresh	48.34	–	x
H. hispanica	499-B	499	1		fresh	48.22	–	x
H. hispanica	499-C	499	1		fresh	48.02*	–	x
H. hispanica	500-11	500	1		dried		–	x
H. hispanica	500-12	500	1		dried		–	x
H. hispanica	500-6	500	1		dried		–	x
H. hispanica	500-7	500	1		dried		–	x
H. hispanica	500-A	500	1		fresh	48.72*	–	x
H. hispanica	500-B	500	1		fresh	49.18*	–	x
H. hispanica	500-C	500	1		fresh	48.16*	–	x
H. hispanica	501-10	501	1		dried		both	x
H. hispanica	501-4	501	1		dried		both	x
H. hispanica	501-5	501	1		dried		both	x
H. hispanica	501-A	501	1		fresh	48.14	both	x
H. hispanica	501-B	501	1		fresh	48.16	both	x
H. hispanica	501-C	501	1		fresh	47.56*	both	x
H. hispanica	501-D	501	1		fresh	48.06*	both	x
H. non-scripta	502-11	502	0		dried		–	x
H. non-scripta	502-13	502	0		dried		–	x
H. non-scripta	502-6	502	0		dried		–	x
H. non-scripta	502-7	502	0		dried		–	x
H. non-scripta	502-A	502	0		fresh	48.34*	–	x
H. non-scripta	502-B	502	NA		fresh	47.68*	–	–
H. non-scripta	502-C	502	0		fresh	48.64*	–	x
H. non-scripta	503-A	503	0		fresh	49.04*	–	x
H. non-scripta	503-B	503	0		fresh	47.96*	–	x
H. non-scripta	503-C	503	0		fresh	48.20*	–	x
H. non-scripta	503-D	503	0		fresh	48.16*	–	x
H. non-scripta	503-E	503	0		fresh	47.91*	–	x
H. non-scripta	503-F	503	0		fresh	47.85*	–	x
H. non-scripta	503-G	503	0		fresh	48.94*	–	x
H. non-scripta	504-10	504	0		dried		–	x
H. non-scripta	504-13	504	0		dried		–	x
H. non-scripta	504-7	504	0		dried		–	x
H. non-scripta	504-8	504	0		dried		–	x
H. non-scripta	504-A	504	0		fresh	2n	–	x
H. non-scripta	504-B	504	NA		fresh	2n	–	–
H. non-scripta	504-C	504	0		fresh		–	x
H. non-scripta	504-D	504	0		fresh		–	x
H. non-scripta	505-10	505	0		dried		–	x
H. non-scripta	505-7	505	0		dried		–	x
H. non-scripta	505-9	505	0		dried		–	x
H. non-scripta	505-A	505	0		fresh	47.74*	–	x
H. non-scripta	505-B	505	0		fresh	47.58*	–	x
H. non-scripta	505-C	505	0		fresh	47.51*	–	x
H. non-scripta	505-D	505	0		fresh	47.16*	–	x
H. non-scripta	506-1	506	0		dried		both	–
H. non-scripta	506-2	506	0		dried		both	–

Table A.1 continued

Species	Individual	Pop	CP	Accession	Tissue	Gs	Chapter 3	Chapter 4
H. non-scripta	506-3	506	0		dried		both	–
H. hispanica	SWA1	339	1	BM000864247	fresh		mRNA	–
H. non-scripta	SWA2	411	0	BM000864320	fresh		mRNA	–
H. paivae	SWA3	130	NA	BM000865089	fresh		mRNA	–
H. hispanica	SWA4	356	NA	BM000864264	fresh		mRNA	–

Table A.2 – List of collecting sites (**ID**) that were included in chapter 2 – chapter 4. The **collector**, collecting **date**, origin (**country/province**) are provided along with **lat**itude, **lon**gitude, and **alt**itude in m above sea level. Each **taxon** is encoded with ns - *H. non-scripta*, hisp - *H. hispanica*, hyN - hybrid North, hyS - hybrid South, and paiv - *H. paivae*. For chapter 2, in which data collection was described, detailed information is provided for the sampled number of **sil**ica material, **bulb**s and whether **pic**tures are available. As collector for new data of this study 'Marquardt et al' is representative, and 'Marquardt et al (revisited)' indicates sites that were previously identified by mostly Grundmann and Rumsey but re-sampled for this study. The number of genome size measurements (**gs**) and flowers flowering in 2014 (**c14**) and 2015 (**c15**) and used in the crossing experiments are also given for each site. For chapter 3 and 4 the number (**N**) of sequenced individuals is given per site. In addition, the origin of the mRNA transcriptome libraries are highlighted (SWA1-4). n.c. – not collected.

								chapter 2							chpt 3	chpt 4
ID	**collector**	**date**	**lat**	**lon**	**alt**	**taxon**	**country/province**	**sil**	**bulb**	**pic**	**m**	**gs**	**c14**	**c15**	**N**	**N**
BB-126	F.J. Rumsey, A.M. Paul	11/04/05	42.87	-7.08	431	ns	Spain/ Galicia				–	–	–	–	2	–
BB-130	F.J. Rumsey, A.M. Paul	12/04/05	42.91	-9.14	NA	paiv	Spain/ Galicia				–	–	–	–	SWA3	–
BB-135	F.J. Rumsey, A.M. Paul	10/04/05	42.86	-7.15	NA	ns	Spain/ Galicia				–	–	–	–	3	3
BB-188	J. Squirrell, M. Grundmann	01/04/08	37.31	-8.55	110	hisp	Portugal/ Faro		1		–	–	–	–	1	–
BB-262	J. Squirrell, M. Grundmann	03/04/08	38.56	-8.93	240	hisp	Portugal/ Setubal				–	–	–	–	2	–
BB-339	M. Grundmann	04/04/08	38.78	-9.45	450	hisp	Spain/ Asturias				–	–	–	–	SWA1	–
BB-346	S.W. Ansell	11/04/08	49.44	1.39	NA	ns	France/ Upper Normandy		1		–	–	–	–	1	–
BB-347	S.W. Ansell	11/04/08	48.87	2.05	NA	ns	France/ Île-de-France		3		–	–	–	–	3	–
BB-353	J.C. Vogel, M. Grundmann	24/04/08	40.07	-8.24	910	hisp	Portugal/ Coimbra		1		–	–	–	–	1	–
BB-356	J.C. Vogel, M. Grundmann	25/04/08	40.41	-7.52	770	hisp	Portugal/ Guarda				–	–	–	–	SWA4	–
BB-361	J.C. Vogel, M. Grundmann	26/04/08	41.96	-6.72	770	hisp	Spain/ Castile and León				–	–	–	–	5	5
BB-391	M. Grundmann, F.J. Rumsey	01/05/08	42.89	-5.76	1120	ns	Spain/ Castile and León				–	–	–	–	5	–
BB-392	M. Grundmann, F.J. Rumsey	01/05/08	42.85	-6.18	1300	ns	Spain/ Castile and León		1		–	–	–	–	1	–
BB-393	M. Grundmann, F.J. Rumsey	01/05/08	42.91	-6.23	1260	ns	Spain/ Castile and León				–	–	–	–	6	5
BB-395	Marquardt et al (revisited)	24/04/13	42.83	-6.51	882	ns	Spain/ Castile and León	9	3	yes	3	1	2	3	–	7
BB-397	M. Grundmann, F.J. Rumsey	02/05/08	42.76	-7.08	825	ns	Spain/ Galicia		2		–	1	–	–	5	5
BB-398	M. Grundmann, F.J. Rumsey	02/05/08	42.69	-7.10	1060	ns	Spain/ Galicia				–	–	–	–	5	5
BB-400	Marquardt et al (revisited)	01/05/13	42.48	-6.56	675	hyN	Spain/ Castile and León	10	3	yes	3	3	2	3	–	7
BB-401	Marquardt et al (revisited)	01/05/13	42.50	-6.52	963	hyN	Spain/ Castile and León	10	4	yes	3	4	1	4	–	7
BB-402	Marquardt et al (revisited)	01/05/13	42.49	-6.48	1148	hyN	Spain/ Castile and León	12	6	yes	4	6	2	4	–	7

Table A.2 continued

ID	collector	date	lat	lon	alt	taxon	country/province	sil	bulb	pic	m	gs	c14	c15	chpt 3	chpt 4
BB-403	Marquardt et al (revisited)	01/05/13	42.50	-6.46	1079	hyN	Spain/ Castile and León	10	5	yes	6	5	6	6	–	7
BB-404	M. Grundmann, F.J. Rumsey	03/05/08	42.71	-6.22	950	ns	Spain/ Castile and León				–	–	–	–	–	5
BB-405	Marquardt et al (revisited)	24/04/13	42.58	-6.44	812	ns	Spain/ Castile and León	10	3	yes	–	3	2	3	–	7
BB-406	Marquardt et al (revisited)	26/04/13	42.46	-6.82	720	ns	Spain/ Castile and León	10	0	yes	–	2	–	–	–	7
BB-407	M. Grundmann, F.J. Rumsey	03/05/08	42.42	-6.66	530	ns	Spain/ Castile and León				–	–	–	–	–	5
BB-410	M. Grundmann, F.J. Rumsey	04/05/08	42.28	-7.46	1010	ns	Spain/ Galicia				–	–	–	–	3	3
BB-411	M. Grundmann, F.J. Rumsey	04/05/08	42.12	-7.95	660	ns	Spain/ Galicia				–	–	–	–	SWA2	–
BB-413	E. Curot-lodeon	03/05/08	47.80	-2.92	NA	ns	France/ Brittany				–	–	–	–	1	–
BB-415	E. Curot-lodeon	03/05/08	48.31	-4.12	NA	ns	France/ Brittany				–	–	–	–	1	–
BB-460	F.J. Rumsey	25/04/09	41.84	-5.98	775	hisp	Spain/ Castile and León				–	–	–	–	6	6
BB-461	F.J. Rumsey	25/04/09	41.88	-6.21	980	hisp	Spain/ Castile and León				–	–	–	–	7	7
BB-462	Marquardt et al (revisited)	29/04/13	42.06	-6.32	853	hisp	Spain/ Castile and León	11	6	yes	1	3	1	3	–	7
BB-472	Marquardt et al	24/04/13	42.70	-6.48	805	ns	Spain/ Castile and León	12	3	yes	3	3	3	2	–	7
BB-473	Marquardt et al	25/04/13	42.48	-6.46	857	hybrids	Spain/ Castile and León	n.c.	n.c.	yes						
BB-474	Marquardt et al	25/04/13	42.66	-6.72	535	ns	Spain/ Castile and León	10	2	yes	1	1	2	1	–	7
BB-475	Marquardt et al	25/04/13	42.64	-6.87	554	ns	Spain/ Castile and León	10	3	yes	–	3	4	2	–	7
BB-476	Marquardt et al	25/04/13	42.64	-6.88	581	ns	Spain/ Castile and León	0	1	yes	–	2	1	1	–	–
BB-477	Marquardt et al	25/04/13	42.63	-6.89	697	ns	Spain/ Castile and León	0	2	yes	3	1	2	1	–	3
BB-478	Marquardt et al	25/04/13	42.57	-6.81	672	ns	Spain/ Castile and León	n.c.	n.c.	no						
BB-479	Marquardt et al	26/04/13	42.41	-6.70	672	hyS	Spain/ Castile and León	10	1	yes	1	1	–	1	–	7
BB-480	Marquardt et al	26/04/13	42.40	-6.63	856	hyS	Spain/ Castile and León	11	3	yes	3	3	3	2	–	7
BB-481	Marquardt et al	26/04/13	42.38	-6.62	971	hyS	Spain/ Castile and León	10	4	yes	2	3	2	2	–	7
BB-482	Marquardt et al	26/04/13	42.35	-6.61	930	hyS	Spain/ Castile and León	12	4	yes	3	3	3	3	–	7
BB-483	Marquardt et al	26/04/13	42.34	-6.52	920	hisp	Spain/ Castile and León	10	8	no	5	5	7	5	–	7
BB-484	Marquardt et al	26/04/13	42.11	-6.73	1009	hisp	Spain/ Castile and León	9	4	no	2	4	–	3	–	7
BB-486	Marquardt et al	27/04/13	42.32	-6.14	910	hyN	Spain/ Castile and León	14	1	yes	1	1	–	1	–	7
BB-487	Marquardt et al	27/04/13	42.38	-6.32	1069	hyN	Spain/ Castile and León	10	1	yes	1	1	–	1	–	7
BB-488	Marquardt et al	27/04/13	42.42	-6.43	1230	hyN	Spain/ Castile and León	5	0	yes	–	3	–	–	–	5
BB-489	Marquardt et al	28/04/13	42.64	-6.27	813	ns	Spain/ Castile and León	10	5	yes	3	2	–	4	–	7

Table A.2 continued

ID	collector	date	lat	lon	alt	taxon	country/province	sil	bulb	pic	m	gs	c14	c15	chpt 3	chpt 4
BB-490	Marquardt et al	29/04/13	42.05	-6.41	947	hisp	Spain/ Castile and León	10	3	yes	2	3	–	2	7	7
BB-491	Marquardt et al	29/04/13	42.03	-6.46	936	hisp	Spain/ Castile and León	10	6	yes	3	3	–	4	–	7
BB-492	Marquardt et al	30/04/13	42.41	-6.69	684	hyS	Spain/ Castile and León	11	5	yes	4	5	3	4	–	6
BB-493	Marquardt et al	30/04/13	42.40	-6.62	1006	hyS	Spain/ Castile and León	11	3	yes	3	1	4	3	–	7
BB-494	Marquardt et al	30/04/13	42.31	-6.49	844	hisp	Spain/ Castile and León	11	3	yes	3	3	3	3	–	7
BB-495	Marquardt et al	01/05/13	42.46	-6.53	1055	hyN	Spain/ Castile and León	10	3	yes	2	3	3	2	–	7
BB-496	Marquardt et al	01/05/13	42.45	-6.55	780	hyN	Spain/ Castile and León	10	2	no	1	2	–	3	–	7
BB-497	Marquardt et al	02/05/13	42.38	-6.64	785	hyS	Spain/ Castile and León	10	6	yes	1	4	–	–	–	7
BB-498	Marquardt et al	02/05/13	42.34	-6.68	866	hyS	Spain/ Castile and León	11	3	yes	3	3	3	2	–	7
BB-499	Marquardt et al	02/05/13	42.26	-6.63	985	hisp	Spain/ Castile and León	14	3	yes	3	3	2	3	–	7
BB-500	Marquardt et al	02/05/13	42.27	-6.59	970	hisp	Spain/ Castile and León	15	3	yes	3	3	–	2	–	7
BB-501	Marquardt et al	02/05/13	42.20	-6.23	945	hisp	Spain/ Castile and León	11	4	yes	3	4	2	2	7	7
BB-502	Marquardt et al	03/05/13	42.55	-6.49	771	ns	Spain/ Castile and León	13	3	yes	2	3	1	2	–	6
BB-503	Marquardt et al	03/05/13	42.54	-6.46	759	ns	Spain/ Castile and León	12	7	yes	5	7	4	4	–	7
BB-504	Marquardt et al	03/05/13	42.65	-6.32	826	ns	Spain/ Castile and León	11	4	yes	3	2	–	1	–	7
BB-505	Marquardt et al	03/05/13	42.88	-6.45	850	ns	Spain/ Castile and León	11	3	yes	4	4	4	4	–	7
BB-506	J. Marquardt	06/04/14	51.22	0.90	NA	ns	United Kingdom/ Kent	3			–	–	–	–	3	–

Bibliography

K. L. Adams and J. D. Palmer. Evolution of mitochondrial gene content: gene loss and transfer to the nucleus. *Molecular Phylogenetics and Evolution*, 29(3):380–395, 2003.

A. A. M. Al-Modayan. *The population dynamics of bluebell (Hyacinthoides non-scripta) in Wayland Wood, Norfolk*. Phd thesis, University of East Anglia, 1993.

S. S. Ali, Y. Yu, M. Pfosser, and W. Wetschnig. Inferences of biogeographical histories within subfamily Hyacinthoideae using S-DIVA and Bayesian binary MCMC analysis implemented in RASP (Reconstruct Ancestral State in Phylogenies). *Annals of Botany*, 109(1):95–107, 2012.

S. S. Ali, M. Pfosser, W. Wetschnig, M. Martinez-Azorin, M. B. Crespo, and Y. Yu. Out of Africa: Miocene dispersal, vicariance, and extinction within Hyacinthaceae subfamily Urgineoideae. *Journal of Integrative Plant Biology*, 55(10):950–64, 2013.

E. C. Anderson and E. A. Thompson. A model-based method for identifying species hybrids using multilocus genetic data. *Genetics*, 160(3):1217–1229, 2002.

S. W. Ansell, N. Bystriakova, H. Schneider, M. Grundmann, F. J. Rumsey, J. C. Vogel, J. Squirrell, and P. Hollingsworth. Tracing the ice age distributions: Integrating climate modelling and phylogeography of iberian bluebells. In prep.

APG II. An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: APG II. *Botanical Journal of the Linnean Society*, 141(4): 399–436, 2003.

APG III. An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: APG III. *Botanical Journal of the Linnean Society*, 161(2): 105–121, 2009.

B. Arnold, R. B. Corbett-Detig, D. Hartl, and K. Bomblies. RADseq underestimates diversity and introduces genealogical biases due to nonrandom haplotype sampling. *Molecular ecology*, 22(11):3179–3190, 2013.

M. L. Arnold. *Natural hybridization and evolution*. Oxford University Press, USA, 1997.

M. L. Arnold and N. H. Martin. Adaptation by introgression. *Journal of Biology*, 8(9): 82, 2009.

M. L. Arnold and N. H. Martin. Hybrid fitness across time and habitats. *Trends in Ecology and Evolution*, 25(9):530–536, 2010.

N. Arrigo, S. Buerki, A. Sarr, R. Guadagnuolo, and G. Kozlowski. Phylogenetics and phylogeography of the monocot genus *Baldellia* (Alismataceae): Mediterranean refugia, suture zones and implications for conservation. *Molecular Phylogenetics and Evolution*, 58(1):33–42, 2011.

E. J. Baack, K. D. Whitney, and L. H. Rieseberg. Hybridization and genome size evolution: timing and magnitude of nuclear DNA content increases in *Helianthus* homoploid hybrid species. *New Phytologist*, 167(2):623–630, 2005.

Babraham Institute, Cambridge, UK. FASTQC: A quality control tool for high throughput sequence data. Technical report, Cambridge, UK: Babraham Institute, 2011.

H. L. Ballard, L. D. Robinson, A. N. Young, G. B. Pauly, L. M. Higgins, R. F. Johnson, and J. C. Tweddle. Contributions to conservation outcomes by natural history museum-led citizen science: Examining evidence and next steps. *Biological Conservation*, 2016.

V. K. Baranwal, V. Mikkilineni, U. B. Zehr, A. K. Tyagi, and S. Kapoor. Heterosis: emerging ideas about hybrid vigour. *Journal of Experimental Botany*, 63(18):6309–6314, 2012.

D. W. Barnett, E. K. Garrison, A. R. Quinlan, M. P. Stromberg, and G. T. Marth. BamTools: a C++ API and toolkit for analyzing and managing BAM files. *Bioinformatics*, 27(12):1691–1692, 2011.

N. H. Barton and K. S. Gale. *Genetic analysis of hybrid zones*, book section 1, pages 13–45. Oxford University Press, 1993.

N. H. Barton and G. M. Hewitt. Analysis of hybrid zones. *Annual review of Ecology and Systematics*, pages 113–148, 1985.

N. H. Barton and G. M. Hewitt. Adaptation, speciation and hybrid zones. *Nature*, 341 (6242):497–503, 1989.

W. Bateson. Heredity and variation in modern lights. *Darwin and modern science*, 85: 101, 1909.

M. A. Beaumont, W. Zhang, and D. J. Balding. Approximate Bayesian computation in population genetics. *Genetics*, 162(4):2025–35, 2002.

J. L. Bella, L. Serrano, J. Orellana, and P. L. Mason. The origin of the *Chorthippus parallelus* hybrid zone: chromosomal evidence of multiple refugia for Iberian populations. *Journal of Evolutionary Biology*, 20(2):568–576, 2007.

K. D. Bennett, P. C. Tzedakis, and K. J. Willis. Quaternary Refugia of North European Trees. *Journal of Biogeography*, 18(1):103–115, 1991.

M. D. Bennett. Nuclear DNA content and minimum generation time in herbaceous plants. *Proceedings of the Royal Society of London Series B, Biological Sciences*, 181(63):109–135, 1972.

M. D. Bennett and J. B. Smith. Nuclear DNA amounts in angiosperms. *Philosophical Transactions of the Royal Society of London Series B, Biological Sciences*, 274(933): 227–74, 1976.

C. W. Birky. Uniparental inheritance of organelle genes. *Current Biology*, 18(16):R692–R695, 2008.

G. E. Blackman and A. J. Rutter. *Endymion non-scriptus* (L.) Garcke. *Journal of Ecology*, 42(2):629–638, 1954.

W. Bleeker, U. Schmitz, and M. Ristow. Interspecific hybridisation between alien and native plant species in Germany and its consequences for native biodiversity. *Biological Conservation*, 137(2):248–253, 2007.

A. M. Bolger, M. Lohse, and B. Usadel. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, 30(15):2114–2120, 2014.

B. M. Bolker, M. E. Brooks, C. J. Clark, S. W. Geange, J. R. Poulsen, M. H. H. Stevens, and J.-S. S. White. Generalized linear mixed models: a practical guide for ecology and evolution. *Trends in Ecology and Evolution*, 24(3):127–135, 2009.

O. Booy, M. Wade, and H. Roy. *Field guide to invasive plants and animals in Britain.* Bloomsbury Publishing, 2015.

A. C. Brennan, G. Woodward, O. Seehausen, V. Muñoz-Fuentes, C. Moritz, A. Guelmami, R. J Abbott, and P. Edelaar. Hybridization due to changing species distributions: adding problems or solutions to conservation of biodiversity during global change? *Evolutionary Ecology Research*, 16(6):475–491, 2015.

S. Brewer, R. Cheddadi, J. L. De Beaulieu, and M. Reille. The spread of deciduous *Quercus* throughout Europe since the last glacial period. *Forest Ecology and Management*, 156 (1):27–48, 2002.

K. J. Brocklebank and G. A. F. Hendry. Characteristics of plant species which store different types of reserve carbohydrates. *New Phytologist*, 112(2):255–260, 1989.

S. Buerki, S. Jose, S. R. Yadav, P. Goldblatt, J. C. Manning, and F. Forest. Contrasting biogeographic and diversification patterns in two Mediterranean-type ecosystems. *PLoS One*, 7(6):e39377, 2012.

R. J. Buggs. Empirical study of hybrid zone movement. *Heredity*, 99(3):301–12, 2007.

C. Buhk and A. Thielsch. Hybridisation boosts the invasion of an alien species complex: Insights into future invasiveness. *Perspectives in Plant Ecology, Evolution and Systematics*, (0), 2015.

M. L. Buide, J. M. Sánchez, and J. Guitián. Ecological characteristics of the flora of the Northwest Iberian Peninsula. *Plant Ecology*, 135(1):1–8, 1998.

R. S. Burton, R. J. Pereira, and F. S. Barreto. Cytonuclear genomic interactions and hybrid breakdown. *Annual Review of Ecology, Evolution, and Systematics*, 44:281–302, 2013.

C. Camacho, G. Coulouris, V. Avagyan, N. Ma, J. Papadopoulos, K. Bealer, and T. L. Madden. Blast+: architecture and applications. *BMC Bioinformatics*, 10(1):421, 2009.

D. R. Campbell, N. M. Waser, and G. T. Pederson. Predicting patterns of mating and potential hybridization from pollinator behavior. *The American Naturalist*, 159(5):438–450, 2002.

M. A. Chapman and R. J. Abbott. Introgression of fitness genes across a ploidy barrier. *New Phytologist*, 186(1):63–71, 2010.

B. Charlesworth and D. Charlesworth. Reproductive isolation: natural selection at work. *Current Biology*, 10(2):R68–R70, 2000.

B. Charlesworth and D. Charlesworth. *Elements of evolutionary genetics*. Roberts Publishers, 2010.

D. Charlesworth. Balancing selection and its effects on sequences in nearby genome regions. *PLoS Genetics*, 2(4):e64, 2006.

M. W. Chase, J. L. Reveal, and M. F. Fay. A subfamilial classification for the expanded asparagalean families Amaryllidaceae, Asparagaceae and Xanthorrhoeaceae. *Botanical Journal of the Linnean Society*, 161(2):132–136, 2009.

S. Chen, D. K. Kim, M. W. Chase, and J. H. Kim. Networks in a large-scale phylogenetic analysis: Reconstructing evolutionary history of Asparagales (Lilianae) based on four plastid genes. *PLoS One*, 8(3), 2013.

S. L. Chown, K. A. Hodgins, P. C. Griffin, J. G. Oakeshott, M. Byrne, and A. A. Hoffmann. Biological invasions, climate change and genomics. *Evolutionary Applications*, 8(1):23–46, 2015.

C. Christe, K. N. Stolting, L. Bresadola, B. Fussi, B. Heinze, D. Wegmann, and C. Lexer. Selection against recombinant hybrids maintains reproductive isolation in hybridizing *Populus* species despite F1 fertility and recurrent gene flow. *Molecular ecology*, 25(11): 2482–2498, 2016.

J. Clark, O. Hidalgo, J. Pellicer, H. Liu, J. Marquardt, Y. Robert, M. Christenhusz, S. Zhang, M. Gibby, I. J. Leitch, and H. Schneider. Genome evolution of ferns: evidence for relative stasis of genome size across the fern phylogeny. *New Phytologist*, 210(3): 1072–82, 2016.

H. P. Comes and J. W. Kadereit. Spatial and temporal patterns in the evolution of the flora of the European alpine system. *Taxon*, 52(3):451–462, 2003.

A. S. Cooke. Effects of grazing by muntjac (*Muntiucus reevesi*) on bluebells (*Hyacinthoides non-scripta*) and a field technique for assessing feeding activity. *Journal of Zoology*, 242 (2):365–369, 1997.

S. A. Corbet. Fruit and seed production in relation to pollination and resources in bluebell, *Hyacinthoides non-scripta*. *Oecologia*, 114(3):349–360, 1998.

R. Cronn, B. J. Knaus, A. Liston, P. J. Maughan, M. Parks, J. V. Syring, and J. Udall. Targeted enrichment strategies for next-generation plant biology. *American Journal of Botany*, 99(2):291–311, 2012.

K. Csilléry, O. François, and M. G. B. Blum. abc: an R package for approximate Bayesian computation (ABC). *Methods in Ecology and Evolution*, 3(3):475–479, 2012.

C. M. Curry. An integrated framework for hybrid zone models. *Evolutionary Biology*, 42 (3):359–365, 2015.

P. Danecek, A. Auton, G. Abecasis, C. A. Albers, E. Banks, M. A. DePristo, R. E. Handsaker, G. Lunter, G. T. Marth, S. T. Sherry, G. McVean, R. Durbin, and Group Genomes Project Analysis. The variant call format and vcftools. *Bioinformatics*, 27 (15):2156–8, 2011.

J. W. Davey and M. L. Blaxter. RADSeq: next-generation population genetics. *Briefings Functional Genomics*, 9(5-6):416–423, 2010.

J. W. Davey, P. A. Hohenlohe, P. D. Etter, J. Q. Boone, J. M. Catchen, and M. L. Blaxter. Genome-wide genetic marker discovery and genotyping using next-generation sequencing. *Nature Reviews Genetics*, 12(7):499–510, 2011.

A. De La Torre, P. K. Ingvarsson, and S. N. Aitken. Genetic architecture and genomic patterns of gene flow between hybridizing species of *Picea*. *Heredity*, 115:153–164, 2015.

A. de Vries and B. D. Ripley. ggdendro: Create dendrograms and tree diagrams using 'ggplot2'. *R package*, 2016.

P. De Wit, M. H. Pespeni, J. T. Ladner, D. J. Barshis, F. Seneca, H. Jaris, N. O. Therkildsen, M. Morikawa, and S. R. Palumbi. The simple fool's guide to population genomics via RNA-Seq: an introduction to high-throughput sequencing data analysis. *Molecular Ecology Resources*, 12(6):1058–1067, 2012.

G. Decocq and M. Hermy. Are there herbaceous dryads in temperate deciduous forests? *Acta botanica gallica*, 150(4):373–382, 2003.

M. A. DePristo, E. Banks, R. Poplin, K. V. Garimella, J. R. Maguire, C. Hartl, A. A. Philippakis, G. del Angel, M. A. Rivas, M. Hanna, A. McKenna, T. J. Fennell, A. M. Kernytsky, A. Y. Sivachenko, K. Cibulskis, S. B. Gabriel, D. Altshuler, and M. J. Daly. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics*, 43(5):491–498, 2011.

J. P. Der, M. S. Barker, N. J. Wickett, C. W. dePamphilis, and P. G. Wolf. *De novo* characterization of the gametophyte transcriptome in bracken fern, *Pteridium aquilinum*. *BMC Genomics*, 12(1):99, 2011.

T. Deroin. Vascular anatomy of the flower of *Hyacinthoides non-scripta* (L.) Chouard ex Rothm. A new insight about a complex placentation pattern in Asparagaceae. *Modern Phytomorphology*, 5:9–18, 2014.

M. Deutsch and M. Long. Intron-exon structures of eukaryotic model organisms. *Nucleic Acids Research*, 27(15):3219–3228, 1999.

A. D'Hont, F. Denoeud, J. M. Aury, F. C. Baurens, F. Carreel, O. Garsmeur, B. Noel, S. Bocs, G. Droc, M. Rouard, C. Da Silva, K. Jabbari, C. Cardi, J. Poulain, M. Souquet, K. Labadie, C. Jourda, J. Lengelle, M. Rodier-Goud, A. Alberti, M. Bernard, M. Correa, S. Ayyampalayam, M. R. McKain, J. Leebens-Mack, D. Burgess, M. Freeling, A. Mbeguie D. Mbeguie, M. Chabannes, T. Wicker, O. Panaud, J. Barbosa, E. Hribova, P. Heslop-Harrison, R. Habas, R. Rivallan, P. Francois, C. Poiron, A. Kilian, D. Burthia, C. Jenny, F. Bakry, S. Brown, V. Guignon, G. Kema, M. Dita, C. Waalwijk, S. Joseph, A. Dievart, O. Jaillon, J. Leclercq, X. Argout, E. Lyons, A. Almeida, M. Jeridi, J. Dolezel, N. Roux, A. M. Risterucci, J. Weissenbach, M. Ruiz, J. C. Glaszmann, F. Quetier, N. Yahiaoui, and P. Wincker. The banana (*Musa acuminata*) genome and the evolution of monocotyledonous plants. *Nature*, 488(7410):213–217, 2012.

I. J. Díaz-Maroto and P. Vila-Lameiro. Deciduous and semi-deciduous oak forests (*Quercus robur*, *Q. petraea* and *Q. pyrenaica*) floristic composition in the Northwest Iberian Peninsula. *Biologia*, 62(2):163–172, 2007.

T. Dines. Stop picking on bluebells. *Botanic Society of the British Isles News*, 98:26–27, 2005.

T. H. Dobzhansky. *Genetics and the Origin of Species*, volume 11. Columbia University Press, 1937.

M. Doebeli and U. Dieckmann. Speciation along environmental gradients. *Nature*, 421 (6920):259–264, 2003.

J. Doležel, J. Greilhuber, and J. Suda. Estimation of nuclear DNA content in plants using flow cytometry. *Nature Protocols*, 2(9):2233–2244, 2007.

J. J. Doyle and J. L. Doyle. A rapid DNA isolation procedure for small quantities of fresh leaf tissue. *Phytochemical Bulletin*, 19:11–15, 1987.

F. Dufresne, M. Stift, R. Vergilino, and B. K. Mable. Recent progress and challenges in population genetics of polyploid organisms: an overview of current state-of-the-art molecular and statistical tools. *Molecular ecology*, 23(1):40–69, 2014.

V. O. Ebuele, A. Santoro, and V. Thoss. Phosphorus speciation by (31)P NMR spectroscopy in bracken (*Pteridium aquilinum* (L.) Kuhn) and bluebell (*Hyacinthoides non-scripta* (L.) Chouard ex Rothm.) dominated semi-natural upland soil. *Science of the Total Environment*, 566-567:1318–1328, 2016.

C. E. Edwards, D. E. Soltis, and P. S. Soltis. Using patterns of genetic structure based on microsatellite loci to test hypotheses of current hybridization, ancient hybridization and incomplete lineage sorting in *Conradina* (Lamiaceae). *Molecular ecology*, 17(23): 5157–5174, 2008.

A. N. Egan, J. Schlueter, and D. M. Spooner. Applications of next-generation sequencing in plant biology. *American Journal of Botany*, 99(2):175–185, 2012.

R. Ekblom and J. Galindo. Applications of next generation sequencing in molecular ecology of non-model organisms. *Heredity*, 107(1):1–15, 2011.

S. El-Metwally, T. Hamza, M. Zakaria, and M. Helmy. Next-generation sequence assembly: four stages of data processing and computational challenges. *PLoS Computational Biology*, 9(12):e1003345, 2013.

H. Ellegren. Genome sequencing and population genomics in non-model organisms. *Trends in Ecology and Evolution*, 29(1):51–63, 2014.

N. C. Ellstrand. Gene flow by pollen: implications for plant conservation genetics. *Oikos*, pages 77–86, 1992.

J. A. Endler. *Geographic variation, speciation, and clines*. Princeton University Press, 1977.

L. Excoffier, P. E. Smouse, and J. M. Quattro. Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data. *Genetics*, 131(2):479–491, 1992.

Y. Fang, H. Wu, T. Zhang, M. Yang, Y. Yin, L. Pan, X. Yu, X. Zhang, S. Hu, I. S. Al-Mssallem, and J. Yu. A complete sequence and transcriptomic analyses of date palm (*Phoenix dactylifera* L.) mitochondrial genome. *PLoS One*, 7(5):e37164, 2012.

G. N. Feliner. Southern European glacial refugia: A tale of tales. *Taxon*, 60(2):365–372, 2011.

N. Ferriday. The field layer of perivale wood. *The London Naturalist*, (66):23–33, 1987.

C. Feuillet, J. E. Leach, J. Rogers, P. S. Schnable, and K. Eversole. Crop genome sequencing: lessons and rationales. *Trends in Plant Science*, 16(2):77–88, 2011.

S. Fischer, B. P. Brunk, F. Chen, X. Gao, O. S. Harb, J. B. Iodice, D. Shanmugam, D. S. Roos, and Jr. Stoeckert, C. J. Using OrthoMCL to assign proteins to OrthoMCL-DB groups or to cluster proteomes into new ortholog groups. *Current Protocols Bioinformatics*, Chapter 6:Unit 6 12 11–19, 2011.

P. Flicek, M. R. Amode, D. Barrell, K. Beal, K. Billis, S. Brent, D. Carvalho-Silva, P. Clapham, G. Coates, S. Fitzgerald, L. Gil, C. G. Giron, L. Gordon, T. Hourlier, S. Hunt, N. Johnson, T. Juettemann, A. K. Kahari, S. Keenan, E. Kulesha, F. J. Martin, T. Maurel, W. M. McLaren, D. N. Murphy, R. Nag, B. Overduin, M. Pignatelli, B. Pritchard, E. Pritchard, H. S. Riat, M. Ruffier, D. Sheppard, K. Taylor, A. Thormann, S. J. Trevanion, A. Vullo, S. P. Wilder, M. Wilson, A. Zadissa, B. L. Aken, E. Birney, F. Cunningham, J. Harrow, J. Herrero, T. J. Hubbard, R. Kinsella, M. Muffato, A. Parker, G. Spudich, A. Yates, D. R. Zerbino, and S. M. Searle. Ensembl 2014. *Nucleic Acids Research*, 42(Database issue):D749–755, 2014.

K. K. Fugate, D. Fajardo, B. Schlautman, J. P. Ferrareze, M. D. Bolton, L. G. Campbell, E. Wiesman, and J. Zalapa. Generation and characterization of a sugarbeet transcriptome and transcript-based SSR markers. *The Plant Genome*, 7(2):0, 2014.

D. Gagliardi and C. J. Leaver. Polyadenylation accelerates the degradation of the mitochondrial mRNA associated with cytoplasmic male sterility in sunflower. *EMBO Journal*, 18(13):3757–3766, 1999.

L. Gay, P. A. Crochet, D. A. Bell, and T. Lenormand. Comparing clines on molecular and phenotypic traits in hybrid zones: a window on tension zone models. *Evolution*, 62(11): 2789–806, 2008.

D. Geerinck. Une épithète pour l'hybride *Hyacinthoides hispanica* (Mill.) Rothm. × *H. non-scripta* (L.) Chouard ex Rothm.: *H.* × *massartiana* Geerinck (Liliaceae). *Belgian Journal of Botany*, 129(1):83–85, 1996.

D. Gilbert. Gene-omes built from mRNAseq not genome DNA. 7th annual Arthropod genomics symposium., 2013.

S. A. Goff, M. Vaughn, S. McKay, E. Lyons, A. E. Stapleton, D. Gessler, N. Matasci, L. Wang, M. Hanlon, A. Lenards, A. Muir, N. Merchant, S. Lowry, S. Mock, M. Helmke, A. Kubach, M. Narro, N. Hopkins, D. Micklos, U. Hilgert, M. Gonzales, C. Jordan, E. Skidmore, R. Dooley, J. Cazes, R. McLay, Z. Lu, S. Pasternak, L. Koesterke, W. H. Piel, R. Grene, C. Noutsos, K. Gendler, X. Feng, C. Tang, M. Lent, S. J. Kim, K. Kvilekval, B. S. Manjunath, V. Tannen, A. Stamatakis, M. Sanderson, S. M. Welch, K. A. Cranston, P. Soltis, D. Soltis, B. O'Meara, C. Ane, T. Brutnell, D. J. Kleibenstein, J. W. White, J. Leebens-Mack, M. J. Donoghue, E. P. Spalding, T. J. Vision, C. R. Myers, D. Lowenthal, B. J. Enquist, B. Boyle, A. Akoglu, G. Andrews, S. Ram, D. Ware, L. Stein, and D. Stanzione. The iplant collaborative: Cyberinfrastructure for plant biology. *Frontiers in Plant Science*, 2:1–16, 2011.

E. Goldberg, K. Kirby, J. Hall, and J. Latham. The ancient woodland concept as a practical conservation tool in Great Britain. *Journal for Nature Conservation*, 15(2): 109–119, 2007.

A. Gómez and D. H. Lunt. *Refugia within refugia: patterns of phylogeographic concordance in the Iberian Peninsula*, book section V, pages 155–188. Springer, 2006.

J. Greilhuber. Intraspecific variation in genome size: a critical reassessment. *Annals of Botany*, 82(suppl 1):27–35, 1998.

S. Greiner and R. Bock. Tuning a ménage à trois: Co-evolution and co-adaptaion of nuclear and organellar genomes in plants. *BioEssays*, 35(4):354–365, 2013.

J. P. Grime, J. G. Hodgson, and R. Hunt. *Comparative plant ecology: a functional approach to common British species*. Springer, 1988.

M. Grundmann, F. J. Rumsey, S. W. Ansell, S. J. Russell, S. C. Darwin, J. C. Vogel, M. Spencer, J. Squirrell, P. M. Hollingsworth, S. Ortiz, and H. Schneider. Phylogeny and taxonomy of the bluebell genus *Hyacinthoides*, Asparagaceae [Hyacinthaceae]. *Taxon*, 59(1):68–82, 2010.

Y. Guo, K. E. Wiegert-Rininger, V. A. Vallejo, C. S. Barry, and R. M. Warner. Transcriptome-enabled marker discovery and mapping of plastochron-related genes in *Petunia* spp. *BMC Genomics*, 16:726, 2015.

B. J. Haas, A. Papanicolaou, M. Yassour, M. Grabherr, P. D. Blood, J. Bowden, M. B. Couger, D. Eccles, B. Li, M. Lieber, M. D. Macmanes, M. Ott, J. Orvis, N. Pochet, F. Strozzi, N. Weeks, R. Westerman, T. William, C. N. Dewey, R. Henschel, R. D. Leduc, N. Friedman, and A. Regev. *De novo* transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nature Protocols*, 8(8): 1494–1512, 2013.

J. A. Hamilton, C. Lexer, and S. N. Aitken. Genomic and phenotypic architecture of a spruce hybrid zone (*Picea sitchensis* x *P. glauca*). *Molecular ecology*, 22(3):827–841, 2013.

J. A. Hamilton, A. R. De la Torre, and S. N. Aitken. Fine-scale environmental variation contributes to introgression in a three-species spruce hybrid complex. *Tree Genetics and Genomes*, 11(1):1–14, 2014.

K. Hanusova, L. Ekrt, P. Vit, F. Kolar, and T. Urfus. Continuous morphological variation correlated with genome size indicates frequent introgressive hybridization among *Diphasiastrum* species (Lycopodiaceae) in Central Europe. *PLoS One*, 9(6):e99552, 2014.

N. Harrison and C. A. Kidner. Next-generation sequencing and systematics: What can a billion base pairs of DNA sequence data do for you? *Taxon*, 60(6), 2011.

R. G. Harrison and E. L. Larson. Heterogeneous genome divergence, differential introgression, and the origin and structure of hybrid zones. *Molecular ecology*, 2016.

E. Haston, J. E. Richardson, P. F. Stevens, M. W. Chase, and D. J. Harris. The Linear Angiosperm Phylogeny Group (LAPG) III: a linear sequence of the families in APG III. *Botanical Journal of the Linnean Society*, 161(2):128–131, 2009.

S. J. Helyar, J. Hemmer-Hansen, D. Bekkevold, M. I. Taylor, R. Ogden, M. T. Limborg, A. Cariani, G. E. Maes, E. Diopere, G. R. Carvalho, and E. E. Nielsen. Application of SNPs for population genetics of nonmodel organisms: new opportunities and challenges. *Molecular Ecology Resources*, 11 Suppl 1:123–136, 2011.

R. Hershberg and D. A. Petrov. Selection on codon bias. *Annual Review of Genetics*, 42: 287–299, 2008.

K. L. Hertweck, M. S. Kinney, S. A. Stuart, O. Maurin, S. Mathews, M. W. Chase, M. A. Gandolfo, and J. C. Pires. Phylogenetics, divergence times and diversification from three genomic partitions in monocots. *Botanical Journal of the Linnean Society*, 178 (3):375–393, 2015.

G. M. Hewitt. Hybrid zones-natural laboratories for evolutionary studies. *Trends in Ecology and Evolution*, 3(7):158–167, 1988.

G. M. Hewitt. Some genetic consequences of ice ages, and their role in divergence and speciation. *Biological Journal of the Linnean Society*, 58(3):247–276, 1996.

G. M. Hewitt. Post-glacial re-colonization of European biota. *Biological Journal of the Linnean Society*, 68(1-2):87–112, 1999.

G. M. Hewitt. Genetic consequences of climatic oscillations in the Quaternary. *Philosophical Transactions of the Royal Society of London Series B, Biological Sciences*, 359 (1442):183–95, 2004.

G. M. Hewitt. Quaternary phylogeography: the roots of hybrid zones. *Genetica*, 139(5): 617–638, 2011.

D. J. Hodkinson and K. Thompson. Plant dispersal: the role of man. *Journal of Applied Ecology*, 34(6):1484–1496, 1997.

J. Hollander, J. Galindo, and R. K. Butlin. Selection on outlier loci and their association with adaptive phenotypes in *Littorina saxatilis* contact zones. *Journal of Evolutionary Biology*, 28(2):328–337, 2015.

I. Höllinger and J. Hermisson. Bounds to parapatric speciation, a Dobzhansky-Muller incompatibilities model involving autosomes, x chromosomes and mitochondria. In prep.

D. J. Howard, G. L. Waring, C. A. Tibbets, and P. G. Gregory. Survival of hybrids in a mosaic hybrid zone. *Evolution*, pages 789–800, 1993.

A. K. Huylmans, A. Lopez Ezquerra, J. Parsch, and M. Cordellier. *De novo* transcriptome assembly and sex-biased gene expression in the cyclical parthenogenetic *Daphnia galeata*. *Genome Biology Evolution*, page evw221, 2016.

J. H. Ietswaart, S. J. M. De Smet, and J. P. M. Lubbers. Hybridization between *Scilla non-scripta* and *S. hispanica* (Liliaceae) in The Netherlands. *Acta Botanica Neerlandica*, 32(5-6):467–480, 1983.

M. Ingrouille. *Historical Ecology of the British Flora*. Chapman and Hall, London, 1995.

M. J. Iriarte-Chiapusso, C. M. Sobrino, L. Gómez-Orellana, B. Hernández-Beloqui, I. García-Moreiras, C. F. Rodriguez, O. Heiri, A. F. Lotter, and P. Ramil-Rego. Reviewing the Lateglacial–Holocene transition in NW Iberia: A palaeoecological approach based on the comparison between dissimilar regions. *Quaternary International*, 403: 211–236, 2016.

J. Josse and F. Husson. missMDA: A package for handling missing values in multivariate data analysis. *Journal of Statistical Software*, 70(1):1–31, 2016.

Z. N. Kamvar, J. F. Tabima, and N. J. Grünwald. Poppr: an R package for genetic analysis of populations with clonal, partially clonal, and/or sexual reproduction. *PeerJ*, 2:e281, 2014.

Z. N. Kamvar, J. C. Brooks, and N. J. Grunwald. Novel R tools for analysis of genome-wide population genetic data with emphasis on clonality. *Frontiers in Genetics*, 6:208, 2015.

A. Kaplunovsky and A. Bolshoy. Statistical analysis of exon lengths in various eukaryotes. *Open access Bioinformatics*, 3:1–15, 2011.

A. Kato, I. Adachi, M. Miyauchi, K. Ikeda, T. Komae, H. Kizu, Y. Kameda, A. A. Watson, R. J. Nash, M. R. Wormald, G. W. J. Fleet, and N. Asano. Polyhydroxylated pyrrolidine and pyrrolizidine alkaloids from *Hyacinthoides non-scripta* and *Scilla campanulata*. *Carbohydrate Research*, 316(1-4):95–103, 1999.

L. J. Kelly and I. J. Leitch. Exploring giant plant genomes with next-generation sequencing technology. *Chromosome Research*, 19(7):939–953, 2011.

A. Khan, I. A. Khan, B. Heinze, and M. K. Azim. The chloroplast genome sequence of date palm (*Phoenix dactylifera* L. cv. 'Aseel'). *Plant Molecular Biology Reporter*, 30(3): 666–678, 2011.

H. Kirk, K. Vrieling, and P. G. Klinkhamer. Maternal effects and heterosis influence the fitness of plant hybrids. *New Phytologist*, 166(2):685–694, 2005.

C. A. Knight, N. A. Molinari, and D. A. Petrov. The large genome constraint hypothesis: evolution, ecology and phenotype. *Annals of Botany*, 95(1):177–190, 2005.

G. H. Knight. Some factors affecting the distribution of *Endymion non-scriptus* (L.) Garcke in Warwickshire Woods. *Journal of Ecology*, 52(2):405–421, 1964.

D. D. Kohn, P. E. Hulme, P. M. Hollingsworth, and A. Butler. Are native bluebells (*Hyacinthoides non-scripta*) at risk from alien congenerics? Evidence from distributions and co-occurrence in Scotland. *Biological Conservation*, 142(1):61–74, 2009.

T. E. Koralewski and K. V. Krutovsky. Evolution of exon-intron structure and alternative splicing. *PLoS One*, 6(3):e18055, 2011.

J. Lachance and S. A. Tishkoff. SNP ascertainment bias in population genetic analyses: why it is important, and how to correct it. *BioEssays*, 35(9):780–786, 2013.

C. Lafon-Placette, M. Vallejo-Marin, C. Parisod, R. J. Abbott, and C. Kohler. Current plant speciation research: unravelling the processes and mechanisms behind the evolution of reproductive isolation barriers. *New Phytologist*, 209(1):29–33, 2016.

S. J. Langdon. Halton bluebell survey report, with updates on future management recommendations. *Chester: rECOrd*, 2007.

B. Langmead and S. L. Salzberg. Fast gapped-read alignment with bowtie 2. *Nature Methods*, 9(4):357–359, 2012.

C. E. Lee. Evolutionary genetics of invasive species. *Trends in Ecology and Evolution*, 17 (8):386–391, 2002.

I. J. Leitch, D. E. Soltis, P. S. Soltis, and M. D. Bennett. Evolution of DNA amounts across land plants (embryophyta). *Annals of Botany*, 95(1):207–217, 2005.

D. A. Levin. The cytoplasmic factor in plant speciation. *Systematic Botany*, 28(1):5–11, 2003.

B. Li, N. Fillmore, Y. Bai, M. Collins, J. A. Thomson, R. Stewart, and C. N. Dewey. Evaluation of *de novo* transcriptome assemblies from RNA-Seq data. *Genome Biol*, 15 (12):553, 2014.

H. Li. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics*, 27(21), 2011.

H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin, and Subgroup Genome Project Data Processing. The sequence alignment/map format and samtools. *Bioinformatics*, 25(16):2078–2079, 2009.

W. Li and A. Godzik. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, 22(13):1658–1659, 2006.

D. Lindtke, C. A. Buerkle, T. Barbara, B. Heinze, S. Castiglione, D. Bartha, and C. Lexer. Recombinant hybrids retain heterozygosity at many loci: new insights into the genomics of reproductive isolation in *Populus*. *Molecular ecology*, 21(20):5042–5058, 2012.

Z. B. Lippman and D. Zamir. Heterosis: revisiting the magic. *Trends in Genetics*, 23(2): 60–66, 2007.

H. E. Lischer and L. Excoffier. PGDSpider: an automated data conversion tool for connecting population genetics and genomics programs. *Bioinformatics*, 28(2):298–299, 2012.

F. C. Ljungqvist. The spatio-temporal pattern of the mid-Holocene thermal maximum. *Geografie–Sborník ČGS*, 116(2):91–110, 2011.

A. J. Lowe and R. J. Abbott. Hybrid swarms: catalysts for multiple evolutionary events in *Senecio* in the British Isles. *Plant Ecology and Diversity*, 8(4):449–463, 2015.

A. Loytynoja and N. Goldman. An algorithm for progressive multiple alignment of sequences with insertions. *Proceedings of the National academy of sciences of the United States of America*, 102(30):10557–10562, 2005.

M. Lucock. Folic acid: nutritional biochemistry, molecular biology, and role in disease processes. *Molecular Genetics and Metabolism*, 71(1-2):121–138, 2000.

J. E. Lunn. Compartmentation in plant metabolism. *Journal of Experimental Botany*, 58 (1):35–47, 2007.

J. C. Manning, P. Goldblatt, and M. F. Fay. A revised generic synopsis of Hyacinthaceae in sub-saharan Africa, based on molecular evidence, including new combinations and the new tribe Pseudoprospereae. *Edinburgh Journal of Botany*, 60(3):533–568, 2003.

N. Mantel. The detection of disease clustering and a generalized regression approach. *Cancer research*, 27(2 Part 1):209–220, 1967.

I. Marques, A. Rosselló-Graell, D. Draper, and J. M. Iriondo. Pollination patterns limit hybridization between two sympatric species of *Narcissus* (Amaryllidaceae). *American Journal of Botany*, 94(8):1352–1359, 2007.

G. Marsaglia, W. W. Tsang, and J. Wang. Evaluating Kolmogorov's distribution. *Journal of Statistical Software*, 8(18), 2003.

M. Martin. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal*, 17(1):10–12, 2011.

W. Martin and R. G. Herrmann. Gene transfer from organelles to the nucleus: how much, what happens, and why? *Plant Physiology*, 118(1):9–17, 1998.

F. Martín-González, L. Barbero, R. Capote, N. Heredia, and G. Gallastegui. Interaction of two successive Alpine deformation fronts: constraints from low-temperature thermochronology and structural mapping (NW Iberian Peninsula). *International Journal of Earth Sciences*, 101(5):1331–1342, 2012.

M. Martinez-Azorin, M. B. Crespo, A. Juan, and M. F. Fay. Molecular phylogenetics of subfamily Ornithogaloideae (Hyacinthaceae) based on nuclear and plastid DNA regions, including a new taxonomic arrangement. *Annals of Botany*, 107(1):1–37, 2011.

G. D. Martinsen, T. G. Whitham, R. J. Turek, and P. Keim. Hybrid populations selectively filter gene introgression between species. *Evolution*, 55(7):1325–1335, 2001.

A. McKenna, M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis, A. Kernytsky, K. Garimella, D. Altshuler, S. Gabriel, M. Daly, and M. A. DePristo. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, 20(9):1297–1303, 2010.

Y. Meng, D. Chen, Y. Jin, C. Mao, P. Wu, and M. Chen. RNA editing of nuclear transcripts in *Arabidopsis thaliana*. *BMC Genomics*, 11 Suppl 4(Suppl 4):S12, 2010.

J. Merryweather and A. Fitter. Arbuscular mycorrhiza and phosphorus as controlling factors in the life history of *Hyacinthoides non-scripta* (L.) Chouard ex Rothm. *New Phytologist*, 129(4):629–636, 1995.

T. P. Michael and S. Jackson. The first 50 plant genomes. *Plant Genome*, 6(2), 2013.

S. G. Michalski and W. Durka. Separation in flowering time contributes to the maintenance of sympatric cryptic plant lineages. *Ecology and evolution*, 5(11):2172–2184, 2015.

M. A. Millar, D. J. Coates, and M. Byrne. Extensive long-distance pollen dispersal and highly outcrossed mating in historically small and disjunct populations of *Acacia woodmaniorum* (Fabaceae), a rare banded iron formation endemic. *Annals of Botany*, page mcu167, 2014.

W. S. Moore. An evaluation of narrow hybrid zones in vertebrates. *Quarterly Review of Biology*, 1977.

J. Moreton, A. Izquierdo, and R. D. Emes. Assembly, assessment, and availability of *de novo* generated eukaryotic transcriptomes. *Frontiers in Genetics*, 6:361, 2015.

D. A. Mulholland, S. L. Schwikkard, and N. R. Crouch. The chemistry and biological activity of the hyacinthaceae. *Natural Product Reports*, 30(9):1165–1210, 2013.

H. J. Muller. Isolating mechanisms, evolution and temperature. *Biological Symposia*, 6: 71–125, 1942.

F. Murtagh and P. Legendre. Ward's hierarchical agglomerative clustering method: Which algorithms implement Ward's criterion? *Journal of Classification*, 31(3):274–295, 2014.

S. Nakagawa and H. Schielzeth. A general and simple method for obtaining R^2 from generalized linear mixed-effects models. *Methods in Ecology and Evolution*, 4(2):133–142, 2013.

K. Nakasugi, R. Crowhurst, J. Bally, and P. Waterhouse. Combining transcriptome assemblies from multiple *de novo* assemblers in the allo-tetraploid plant *Nicotiana benthamiana*. *PLoS One*, 9(3):e91776, 2014.

M. Nei and R. K. Chesser. Estimation of fixation indices and gene diversities. *Annals of Human Genetics*, 47(Pt 3):253–259, 1983.

J. A. Nelder and R. W. M. Wedderburn. Generalized linear models. *Journal of the Royal Statistical Society. Series A (General)*, 135(3):370–384, 1972.

R. W. Ness, M. Siol, and S. C. Barrett. *De novo* sequence assembly and characterization of the floral transcriptome in cross- and self-fertilizing plants. *BMC Genomics*, 12(1): 298, 2011.

A. W. Nolte, Z. Gompert, and C. A. Buerkle. Variable patterns of introgression in two sculpin hybrid zones suggest that genomic isolation differs among populations. *Molecular ecology*, 18(12):2615–2627, 2009.

P. Nosil. Speciation with gene flow could be common. *Molecular ecology*, 17(9):2103–2106, 2008.

R. Ophir, A. Sherman, M. Rubinstein, R. Eshed, M. Sharabi Schwager, R. Harel-Beja, I. Bar-Ya'akov, and D. Holland. Single-nucleotide polymorphism markers from *de novo* assembly of the pomegranate transcriptome reveal germplasm genetic diversity. *PLoS One*, 9(2):e88998, 2014.

H. J. O'Regan. The Iberian Peninsula - corridor or cul-de-sac? Mammalian faunal change and possible routes of dispersal in the last 2 million years. *Quaternary Science Reviews*, 27(23-24):2136–2144, 2008.

H. A. Orr. Dobzhansky, Bateson, and the genetics of speciation. *Genetics*, 144(4):1331–1335, 1996.

S. Ortiz and J. Rodríguez-Oubiña. Taxonomic characterization of populations of *Hyacinthoides* sect. Somera (Hyacinthaceae) in the northwestern Iberian Peninsula. *Plant Systematics and Evolution*, 202(1-2):111–119, 1996.

S. Ortiz, M. Buján, and J. Rodríguez-Oubiña. A revision of European taxa of *Hyacinthoides* section Somera (Hyacinthaceae) on the basis of multivariate analysis. *Plant Systematics and Evolution*, 217(1-2):163–175, 1999.

K. W. Page. Hybrid bluebells. *Botanical Society of the British Isles News*, 46:9, 1987.

C. Palma-Silva, M. Ferro, M. Bacci, and A. C. Turchetto-Zolet. *De novo* assembly and characterization of leaf and floral transcriptomes of the hybridizing bromeliad species (*Pitcairnia* spp.) adapted to Neotropical Inselbergs. *Molecular Ecology Resources*, 2016.

A. S. Papadopulos, W. J. Baker, D. Crayn, R. K. Butlin, R. G. Kynast, I. Hutton, and V. Savolainen. Speciation with gene flow on Lord Howe Island. *Proceedings of the National Academy of Sciences*, 108(32):13188–13193, 2011.

R. J. Petit, I. Aguinagalde, J.-L. de Beaulieu, C. Bittkau, S. Brewer, R. Cheddadi, R. Ennos, S. Fineschi, D. Grivet, and M. Lascoux. Glacial refugia: hotspots but not melting pots of genetic diversity. *Science*, 300(5625):1563–1565, 2003.

M. Pfosser and F. Speta. Phylogenetics of Hyacinthaceae based on plastid DNA sequences. *Annals of the Missouri Botanical Garden*, 86(4):852–875, 1999.

M. Pfosser, W. Wetschnig, S. Ungar, and G. Prenner. Phylogenetic relationships among genera of Massonieae (Hyacinthaceae) inferred from plastid DNA and seed morphology. *Journal of Plant Research*, 116(2):115–132, 2003.

C. D. Pigott. The flora and vegetation of Britain - ecology and conservation. *New Phytologist*, 98(1):119–128, 1984.

E. Pilgrim and N. Hutchinson. Bluebells for britain – the report of the 2003 bluebells for britain survey. Report, Plantlife International, 2004.

J. Prancl, Z. Kaplan, P. Travnicek, and V. Jarolimova. Genome size as a key to evolutionary complex aquatic plants: polyploidy and hybridization in *Callitriche* (Plantaginaceae). *PLoS One*, 9(9):e105997, 2014.

C. D. Preston, D. A. Pearman, and T. D. Dines. *New atlas of the British and Irish flora: an atlas of the vascular plants of Britain, Ireland, the Isle of Man and the Channel Islands*. Oxford University Press, Oxford, 2002.

H. J. Price, G. Hodnett, and J. S. Johnston. Sunflower (*Helianthus annuus*) leaves contain compounds that reduce nuclear propidium iodide fluorescence. *Annals of Botany*, 86 (5):929–934, 2000.

W. Qi, F. Lin, Y. Liu, B. Huang, J. Cheng, W. Zhang, and H. Zhao. High-throughput development of simple sequence repeat markers for genetic diversity research in *Crambe abyssinica*. *BMC Plant Biology*, 16(1):139, 2016.

A. J. Quené-Boterenbrood. Over het voorkomen van *Scilla non-scripta* (L.) Hoffmanns. and Link, *S. hispanica* Miller en hun hybride in Nederland. *Gorteria: tijdschrift voor de floristiek, de plantenoecologie en het vegetatie-onderzoek van Nederland*, 12(5):91–104, 1984.

A. R. Quinlan and I. M. Hall. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26(6):841–842, 2010.

R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. 2016, *www.R-project.org*. 2016.

O. Rackham and A.T. Grove. *The nature of Mediterranean Europe: an ecological history*. Yale University Press, New Haven, CT, 2001.

A. Raj, M. Stephens, and J. K. Pritchard. fastSTRUCTURE: variational inference of population structure in large SNP data sets. *Genetics*, 197(2):573–589, 2014.

P. Ramil-Rego, C. Muñoz-Sobrino, M. Rodríguez-Guitián, and L. Gómez-Orellana. Differences in the vegetation of the North Iberian Peninsula during the last 16,000 years. *Plant Ecology*, 138(1):41–62, 1998.

J. Ramsey and D. W. Schemske. Pathways, mechanisms, and rates of polyploid formation in flowering plants. *Annual review of Ecology and Systematics*, pages 467–501, 1998.

D. M. Rand and R. G. Harrison. Ecological genetics of a mosaic hybrid zone: mitochondrial, nuclear, and reproductive differentiation of crickets by soil type. *Evolution*, pages 432–449, 1989.

F. Rebeille, C. Alban, J. Bourguignon, S. Ravanel, and R. Douce. The role of plant mitochondria in the biosynthesis of coenzymes. *Photosynth Res*, 92(2):149–162, 2007.

E. Recuero and M. Garcia-Paris. Evolutionary history of *Lissotriton helveticus*: multilocus assessment of ancestral vs. recent colonization of the Iberian Peninsula. *Molecular Phylogenetics and Evolution*, 60(1):170–182, 2011.

A. S. Reddy, Y. Marquez, M. Kalyna, and A. Barta. Complexity of the alternative splicing landscape in plants. *Plant Cell*, 25(10):3657–3683, 2013.

N. Remon, P. Galan, M. Vila, O. Arribas, and H. Naveira. Causes and evolutionary consequences of population subdivision of an Iberian mountain lizard, *Iberolacerta monticola*. *PLoS One*, 8(6):e66034, 2013.

R. W. Riding. White flowered bluebells (*Endymion non-scriptus* (L.) Garcke). *Watsonia*, 11:255, 1977.

L. H. Rieseberg. Crossing relationships among ancient and experimental sunflower hybrid lineages. *Evolution*, 54(3):859–865, 2000.

F. Rose. Indicators of ancient woodland-the use of vascular plants in evaluating ancient woods for nature conservation. *British Wildlife*, 10:241–251, 1999.

T. Sahr, S. Ravanel, and F. Rebeille. Tetrahydrofolate biosynthesis and distribution in higher plants. *Biochem Soc Trans*, 33(Pt 4):758–762, 2005.

L. R. Salgado, D. M. Koop, D. G. Pinheiro, R. Rivallan, V. Le Guen, M. F. Nicolas, L. G. de Almeida, V. R. Rocha, M. Magalhaes, A. L. Gerber, A. Figueira, J. C. Cascardo, A. R. de Vasconcelos, Jr. Silva, W. A., L. L. Coutinho, and D. Garcia. *De novo* transcriptome analysis of *Hevea brasiliensis* tissues by RNA-seq and screening for molecular markers. *BMC Genomics*, 15(1):236, 2014.

J. B. Sambatti, D. Ortiz-Barrientos, E. J. Baack, and L. H. Rieseberg. Ecological selection maintains cytonuclear incompatibilities in hybridizing sunflowers. *Ecology Letters*, 11 (10):1082–1091, 2008.

G. Sarah, F. Homa, S. Pointet, S. Contreras, F. Sabot, B. Nabholz, S. Santoni, L. Saune, M. Ardisson, N. Chantret, C. Sauvage, J. Tregear, C. Jourda, D. Pot, Y. Vigouroux, H. Chair, N. Scarcelli, C. Billot, N. Yahiaoui, R. Bacilieri, B. Khadari, M. Boccara, A. Barnaud, J. P. Peros, J. P. Labouisse, J. L. Pham, J. David, S. Glemin, and M. Ruiz. A large set of 26 new reference transcriptomes dedicated to comparative population genomics in crops and wild relatives. *Molecular Ecology Resources*, 2016.

J. F. Scheepens, E. S. Frei, G. F. J. Armbruster, and J. Stöcklin. Pollen dispersal and gene flow within and into a population of the alpine monocarpic plant *Campanula thyrsoides*. *Annals of Botany*, 110(7):1479–1488, 2012.

K. A. Schierenbeck and N. C. Ellstrand. Hybridization and the evolution of invasiveness in plants and other organisms. *Biological Invasions*, 11(5):1093–1105, 2009.

S. Schliesky, U. Gowik, A. P. Weber, and A. Brautigam. RNA-Seq assembly - Are we there yet? *Frontiers in Plant Science*, 3:220, 2012.

C. Schlötterer. The evolution of molecular markers—just a matter of fashion? *Nature Reviews Genetics*, 5(1):63–69, 2004.

T. Schmitt. Molecular biogeography of europe: Pleistocene cycles and postglacial trends. *Frontiers in Zoology*, 4:11, 2007.

H. Schneider, S. J. Russell, C. J. Cox, F. Bakker, S. Henderson, F. J. Rumsey, J. A. Barrett, M. Gibby, and J. C. Vogel. Chloroplast phylogeny of asplenioid ferns based on *trnL-F* spacer sequences (Polypodiidae, Aspleniaceae) and its implications for biogeography. *Systematic Botany*, 29(2):260–274, 2004.

C. N. Schoebel, S. Brodbeck, D. Buehler, C. Cornejo, J. Gajurel, H. Hartikainen, D. Keller, M. Leys, S. Ricanova, G. Segelbacher, S. Werth, and D. Csencsics. Lessons learned from microsatellite development for nonmodel organisms using 454 pyrosequencing. *Journal of Evolutionary Biology*, 26(3):600–611, 2013.

M. H. Schulz, D. R. Zerbino, M. Vingron, and E. Birney. Oases: robust *de novo* RNA-seq assembly across the dynamic range of expression levels. *Bioinformatics*, 28(8):1086–1092, 2012.

K. Schwenk, N. Brede, and B. Streit. Introduction. Extent, processes and evolutionary impact of interspecific hybridization in animals. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 363(1505):2805–2811, 2008.

B. B. Sears. Elimination of plastids during spermatogenesis and fertilization in the plant kingdom. *Plasmid*, 4(3):233–55, 1980.

A. Sedghifar, Y. Brandvain, P. Ralph, and G. Coop. The spatial mixing of genomes in secondary contact zones. *Genetics*, 201(1):243–261, 2015.

J. E. Seeb, G. Carvalho, L. Hauser, K. Naish, S. Roberts, and L. W. Seeb. Single-nucleotide polymorphism (SNP) discovery and applications of SNP genotyping in nonmodel organisms. *Molecular Ecology Resources*, 11 (Suppl 1):1–8, 2011.

E. Serrano, J. J. González-Trueba, R. Pellitero, and M. Gómez-Lende. Quaternary glacial history of the Cantabrian Mountains of northern Spain: a new synthesis. *Geological Society, London, Special Publications*, 433:SP433. 438, 2016.

H. Sheehan, K. Hermann, and C. Kuhlemeier. Color and scent: how single genes influence pollinator attraction. In *Cold Spring Harbor symposia on quantitative biology*, volume 77, pages 117–133. Cold Spring Harbor Laboratory Press, 2012.

G. H. Shull. The composition of a field of maize. *Journal of Heredity*, (1):296–301, 1908.

B. P. Shumaker and R. W. Sinnott. Astronomical computing: 1. Computing under the open sky. 2. Virtues of the haversine. *Sky and telescope*, 68:158–159, 1984.

M. Simmonds. DNA and phytochemistry of bluebells. *Curtis's Botanical Magazine*, 21 (1):103–104, 2004.

N. K. Sims, E. A. John, and A. J. A. Stewart. Short-term response and recovery of bluebells (*Hyacinthoides non-scripta*) after rooting by wild boar (*Sus scrofa*). *Plant Ecology*, 215(12):1409–1416, 2014.

A. Sindhu, L. Ramsay, L. A. Sanderson, R. Stonehouse, R. Li, J. Condie, A. S. Shunmugam, Y. Liu, A. B. Jha, M. Diapari, J. Burstin, G. Aubert, B. Tar'an, K. E. Bett, T. D. Warkentin, and A. G. Sharpe. Gene-based SNP discovery and genetic mapping in pea. *Theoretical and Applied Genetics*, 127(10):2225–2241, 2014.

E. A. Slade and D. R. Causton. The germination of some woodland herbaceous species under laboratory conditions: a multifactorial study. *New Phytologist*, 83(2):549–557, 1979.

G. S. Slater and E. Birney. Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics*, 6(1):31, 2005.

C. M. Sobrino, P. Ramil-Rego, and L. Gómez-Orellana. Late Würm and early Holocene in the mountains of northwest Iberia: biostratigraphy, chronology and tree colonization. *Vegetation History and Archaeobotany*, 16(4):223–240, 2007.

D. E. Soltis, M. A. Gitzendanner, G. Stull, M. Chester, A. Chanderbali, S. Chamala, I. Jordon-Thaden, P. S. Soltis, P. S. Schnable, and W. B. Barbazuk. The potential of genomics in plant systematics. *Taxon*, 62(5):886–898, 2013.

P. R. Staab, S. Zhu, D. Metzler, and G. Lunter. scrm: efficiently simulating long sequences using the approximated coalescent with recombination. *Bioinformatics*, page btu861, 2015.

C. A. Stace. *New flora of the British Isles*. Cambridge University Press, Cambridge, U.K., second edition, 1997.

S. Stankowski, J. M. Sobel, and M. A. Streisfeld. Geographic cline analysis as a tool for studying genome-wide variation: a case study of pollinator-mediated divergence in a monkeyflower. *Molecular ecology*, page 036954, 2016.

P. R. Steele, K. L. Hertweck, D. Mayfield, M. R. McKain, J. Leebens-Mack, and J. C. Pires. Quality and quantity of data recovered from massively parallel sequencing: Examples in Asparagales and Poaceae. *American Journal of Botany*, 99(2):330–348, 2012.

R. G. Stickland and B. J. Harrison. Precursors and genetic control of pigmentation 3. Detection and distribution of different white genotypes of bluebells (*Endymion* species). *Heredity*, 39(3):327–333, 1977.

J. C Stout. Plant invasions: Their threats in an Irish context. In *Biology and Environment: Proceedings of the Royal Irish Academy*, pages 135–141. JSTOR, 2011.

S. R. Strickler, A. Bombarely, and L. A. Mueller. Designing a transcriptome next-generation sequencing project for a nonmodel plant species. *American Journal of Botany*, 99(2):257–266, 2012.

J.-C. Svenning and F. Skov. Could the tree diversity pattern in Europe be generated by postglacial dispersal limitation? *Ecology letters*, 10(6):453–460, 2007a.

J.-C. Svenning and F. Skov. Ice age legacies in the geographical distribution of tree species richness in Europe. *Global Ecology and Biogeography*, 16(2):234–245, 2007b.

P. Taberlet, L. Fumagalli, A. G. Wust-Saucy, and J. F. Cosson. Comparative phylogeography and postglacial colonization routes in Europe. *Mol Ecol*, 7(4):453–64, 1998.

P. Tarroso, J. Carrión, M. Dorado-Valiño, P. Queiroz, L. Santos, A. Valdeolmillos-Rodríguez, P. Célio Alves, J. C. Brito, and R. Cheddadi. Spatial climate dynamics in the Iberian Peninsula since 15 000 yr BP. *Climate of the Past*, 12(5):1137–1149, 2016.

K. C. Teeter, L. M. Thibodeau, Z. Gompert, C. A. Buerkle, M. W. Nachman, and P. K. Tucker. The variable genomic architecture of isolation between hybridizing species of house mice. *Evolution*, 64(2):472–485, 2010.

Thermo Scientific. Assessment of nucleic acid purity. *T042-Technical Bulletin Nano Drop Spectrophotometers*, 2013.

J. Thioulouse, D. Chessel, S. Dole, and J.-M. Olivier. ADE-4: a multivariate analysis and graphical display software. *Statistics and computing*, 7(1):75–83, 1997.

P. A. Thompson and S. A. Cox. Germination of bluebell (*Hyacinthoides non-scripta* (L.) Chouard) in relation to its distribution and habitat. *Annals of Botany*, 42(177), 1978.

V. Thoss, P. J. Murphy, R. Marriott, and T. Wilson. Triacylglycerol composition of British bluebell (*Hyacinthoides non-scripta*) seed oil. *RSC Advances*, 2(12):5314–5322, 2012.

M. Todesco, M. A. Pascual, G. L. Owens, K. L. Ostevik, B. T. Moyers, S. Hübner, S. M. Heredia, M. A. Hahn, C. Caseys, and D. G. Bock. Hybridization and extinction. *Evolutionary Applications*, 2016.

S. A. Trewick, M. Morgan-Richards, S. J. Russell, S. Henderson, F. J. Rumsey, I. Pinter, J. A. Barrett, M. Gibby, and J. C. Vogel. Polyploidy, phylogeography and Pleistocene refugia of the rockfern *Asplenium ceterach*: evidence from chloroplast DNA. *Molecular ecology*, 11(10):2003–2012, 2002.

W. B. Turrill. *Endymion hispanicus*, Curtis's Bot. *Mag., NS*, 176, 1952.

A. D. Twyford and R. A. Ennos. Next-generation hybridization and introgression. *Heredity*, 108(3):179–189, 2011.

A. Untergasser, I. Cutcutache, T. Koressaar, J. Ye, B. C. Faircloth, M. Remm, and S. G. Rozen. Primer3–new capabilities and interfaces. *Nucleic Acids Research*, 40(15):e115–e115

S. Uribe-Convers, M. L. Settles, and D. C. Tank. A phylogenomic approach based on PCR target enrichment and high throughput sequencing: resolving the diversity within the South American species of *Bartsia* L.(Orobanchaceae). *PLoS One*, 11(2):e0148203, 2016.

J. P. Vähä and C. R. Primmer. Efficiency of model-based bayesian methods for detecting hybrid individuals under different hybridization scenarios and with different numbers of loci. *Molecular ecology*, 15(1):63–72, 2006.

G. A. Van der Auwera, M. O. Carneiro, C. Hartl, R. Poplin, G. Del Angel, A. Levy-Moonshine, T. Jordan, K. Shakir, D. Roazen, J. Thibault, E. Banks, K. V. Garimella, D. Altshuler, S. Gabriel, and M. A. DePristo. From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. *Current Protocols Bioinformatics*, 43:11.10.1–11.10.33, 2013.

S. van der Veken, J. Rogister, K. Verheyen, M. Hermy, and R. A. N. Nathan. Over the (range) edge: a 45-year transplant experiment with the perennial forest herb *Hyacinthoides non-scripta*. *Journal of Ecology*, 95(2):343–351, 2007.

L. van Herwerden, J. H. Choat, C. L. Dudgeon, G. Carlos, S. J. Newman, A. Frisch, and M. van Oppen. Contrasting patterns of genetic structure in two species of the coral trout *Plectropomus* (Serranidae) from east and west Australia: introgressive hybridisation or ancestral polymorphisms. *Molecular Phylogenetics and Evolution*, 41(2):420–435, 2006.

F. Vandelook and J. A. van Assche. Temperature requirements for seed germination and seedling development determine timing of seedling emergence of three monocotyledonous temperate forest spring geophytes. *Annals of Botany*, 102(5):865–875, 2008.

M. Vatanparast, P. Shetty, R. Chopra, J. J. Doyle, N. Sathyanarayana, and A. N. Egan. Transcriptome sequencing and marker development in winged bean (*Psophocarpus tetragonolobus*; Leguminosae). *Sci Rep*, 6:29070, 2016.

Y. Vega, I. Marques, S. Castro, and J. Loureiro. Outcomes of extensive hybridization and introgression in *Epidendrum* (Orchidaceae): Can we rely on species boundaries? *PLoS One*, 8(11):e80662, 2013.

E. Vendramin, M. T. Dettori, J. Giovinazzi, S. Micali, R. Quarta, and I. Verde. A set of EST-SSRs isolated from peach fruit transcriptome and their transportability across *Prunus* species. *Molecular Ecology Notes*, 7(2):307–310, 2007.

S. Via. Divergence hitchhiking and the spread of genomic isolation during ecological speciation-with-gene-flow. *Philosophical Transactions of the Royal Society of London Series B, Biological Sciences*, 367(1587):451–460, 2012.

A. E. Vinogradov. Selfish DNA is maladaptive: evidence from the plant red list. *Trends in Genetics*, 19(11):609–614, 2003.

E. A. Visser, J. L. Wegrzyn, E. T. Steenkmap, A. A. Myburg, and S. Naidoo. Combined *de novo* and genome guided assembly and annotation of the *Pinus patula* juvenile shoot transcriptome. *BMC Genomics*, 16(1):1057, 2015.

J. Walsh, W. G. Shriver, B. J. Olsen, and A. I. Kovach. Differential introgression and the maintenance of species boundaries in an advanced generation avian hybrid zone. *BMC Evolutionary Biology*, 16(1):65, 2016.

G. H. Walter and R. Hengeveld. The structure of the two ecological paradigms. *Acta Biotheoretica*, 48(1):15–46, 2000.

R. X. Wang. Gene flow across a hybrid zone maintained by a weak heterogametic incompatibility and positive selection of incompatible alleles. *Journal of Evolutionary Biology*, 26(2):386–398, 2013.

H. Wanner, L. Mercolli, M. Grosjean, and S. P. Ritz. Holocene climate variability and change; a data-based review. *Journal of the Geological Society*, 172(2):254–263, 2015.

M. Ward, S. D. Johnson, and M. P. Zalucki. Modes of reproduction in three invasive milkweeds are consistent with baker's rule. *Biological Invasions*, 14(6):1237–1250, 2012.

J. H. Ward Jr. Hierarchical grouping to optimize an objective function. *Journal of the American statistical association*, 58(301):236–244, 1963.

A. A. Watson, R. J. Nash, M. R. Wormald, D. J. Harvey, S. Dealler, E. Lees, N. Asano, H. Kizu, A. Kato, R. C. Griffiths, A. J. Cairns, and G. W. J. Fleet. Glycosidase-inhibiting pyrrolidine alkaloids from *Hyacinthoides non-scripta*. *Phytochemistry*, 46(2): 255–259, 1997.

W. Wetschnig and M. Pfosser. The *Scilla plumbea* puzzle-present status of the genus *Scilla sensu lato* in southern Africa and description of *Spetaea lachenaliiflora*, a new genus and species of Massonieae (Hyacinthaceae). *Taxon*, 52(1):75–91, 2003.

K. D. Whitney, J. R. Ahern, L. G. Campbell, L. P Albert, and M. S King. Patterns of hybridization in plants. *Perspectives in Plant Ecology, Evolution and Systematics*, 12 (3):175–182, 2010.

P. Willmer. *Pollination and floral ecology*. Princeton University Press, 2011.

J. Y. Wilson. Polyploidy in bluebells (*Endymion non-scriptus* and *E. hispanicus*). *Nature*, 178(4526):195–196, 1956.

J. Y. Wilson. Cytogenetics of triploid bluebells *Endymion nonscriptus* (L.) Garcke and *E. hispanicus* (Mill.) Chouard. *Cytologia*, 23(4):435–446, 1958.

J. Y. Wilson. Vegetative reproduction in the bluebell, *Endymion nonscriptus* (L.) Garcke. *New Phytologist*, 58(2):155–163, 1959a.

J. Y. Wilson. Verification of the breeding system in the bluebell *Endymion non-scriptus* (L.) Garcke. *Annals of Botany*, 33(89), 1959b.

D. E. Wolf, N. Takebayashi, and L. H. Rieseberg. Predicting the risk of extinction through hybridization. *Conservation Biology*, 15(4):1039–1053, 2001.

Inc. Wolfram Research. Mathematica, 2016.

T. W. Woodhead. Notes on the bluebell. *The Naturalist*, 565, 41–48, and 566, 81–88, 1904.

S. Wright. *Variability within and among natural populations*, volume 4 of *Evolution and the Genetics of Populations*. University of Chicago Press, 1978.

Y. Xie, G. Wu, J. Tang, R. Luo, J. Patterson, S. Liu, W. Huang, G. He, S. Gu, S. Li, X. Zhou, T. W. Lam, Y. Li, X. Xu, G. K. Wong, and J. Wang. SOAPdenovo-Trans: *de novo* transcriptome assembly with short RNA-Seq reads. *Bioinformatics*, pages 1–7, 2014.

M. Yang, X. Zhang, G. Liu, Y. Yin, K. Chen, Q. Yun, D. Zhao, I. S. Al-Mssallem, and J. Yu. The complete chloroplast genome sequence of date palm (*Phoenix dactylifera* L.). *PLoS One*, 5(9):e12762, 2010.

P. S. Yustos and F. D. Martín. Dancing to the rhythms of the Pleistocene? Early Middle Paleolithic population dynamics in NW Iberia (Duero Basin and Cantabrian Region). *Quaternary Science Reviews*, 121:75–88, 2015.

W. H. Zagwijn. Migration of vegetation during the Quaternary in Europe. *Courier Forschungsinstitut Senckenberg*, 153:9–20, 1992.

D. R. Zerbino and E. Birney. Velvet: algorithms for *de novo* short read assembly using de Bruijn graphs. *Genome Research*, 18(5):821–829, 2008.

L. Zhu, Y. Zhang, W. Zhang, S. Yang, J. Q. Chen, and D. Tian. Patterns of exon-intron architecture variation of genes in eukaryotic genomes. *BMC Genomics*, 10(1):47, 2009.

J. Zohren, N. Wang, I. Kardailsky, J. S. Borrell, A. Joecker, R. A. Nichols, and R. J. A. Buggs. Unidirectional diploid-tetraploid introgression among British birch trees with shifting ranges shown by RAD markers. *Molecular ecology*, 2016.

A. F. Zuur, E. N. Ieno, and A. A. Saveliev. *Mixed Effects Models and Extensions in Ecology with R*. Springer, 2009.